National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

NL-339 (r. 88/04) c

Canada

National Library of Canada
Canadian Theses Service

Bibliothèque nationale du Canada
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-55644-7

Canada

THE UNIVERSITY OF ALBERTA

RESPONSE MODE EFFECTS IN SUBJECTIVE PROBABILITY ASSESSMENT: EXPERIMENTS IN AUDITING JUDGMENT

by Craig Emby

(C)

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
ACCOUNTING

FACULTY OF BUSINESS

EDMONTON, ALBERTA
(FALL) (1989)

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Craig E.N. Emby

NAME OF THESIS: Response Mode Effects in Subjective Probability Assessment: Experiments in Auditing Judgment

DEGREE: Doctor of Philosophy

YEAR THIS DEGREE GRANTED: 1989

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Student's signature)

1640 Nanaimo St.
New Westminster, B.C.

(Student's permanent address)

Date: October 2, 1989.

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled: RESPONSE MODE EFFECTS IN SUBJECTIVE PROBABILITY ASSESSMENT: EXPERIMENTS IN AUDITING JUDGMENT

submitted by: CRAIG E.N. EMBY
in partial fulfilment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in ACCOUNTING

....
(Supervisor)

W R Scott
....

....

J D Jobson
....

Brendan G Rule
....

....

Date: October 4, 1989

ABSTRACT

This study reports the results of two experiments which examined auditors' judgment. The environment in which this judgment must be exercised is characterized by uncertainty. The Probabilistic Judgment Paradigm provides a methodology for decision making which explicitly recognizes this uncertainty. This study presents the results of two experiments, one involving Fraud Investigation and one concerning Internal Control Risk Evaluation, which required the auditor-subjects to make subjective probability judgments based on indirect, qualitative evidence. The Fraud Investigation experiment examined the effects of providing the auditors with different response modes to express their judgments, and differing levels of prior tutoring in axiomatic probability theory. The Internal Control Risk Evaluation experiment also examined the effects of response mode and prior training on probabilistic decision making. In addition, it compared the effect of the two probabilistic modes of judgment with two non-probabilistic modes, on subsequent judgments concerning appropriate level of substantive testing. A visually-based graphical response mode which allowed the auditors to make their probability judgments without the need for complex arithmetical processing eliminated both the violations of probability axioms and the non-conformance with Bayesian patterns of belief revision, with or without any prior tutoring in formal probability theory. The visually-based graphical response mode also enhanced secondary characteristics of the judgments such as consensus, confidence and speed of decision making, with respect to the judgments themselves (Fraud Investigation experiment) and with respect to subsequent decisions based on those judgments (Internal Control Risk Evaluation experiment).

TABLE OF CONTENTS

CHAPTER PAGE

I. INTRODUCTION 1

Outline of the paper 3

II. MOTIVATION AND CONTRIBUTION 4

Summary 8

III. PROBABILITY AND DECISION MAKING UNDER UNCERTAINTY 10

(a) Introduction 10

(b) Multiplicative Probability 11

(c) Additive Probability 13

(d) Cumulative Probability 14

(e) Empirical Results 14

Summary 15

IV. JUDGMENT IN AUDITING 16

(a) Introduction 16

(b) Background 16

(c) Probabilistic Judgment 18

(i) Probability elicitation 18

(ii) Probability revision 21

(iii) Probability manipulation 22

(iv) Summary of probability studies 24

(d) Graphical Decision Aids 24

Summary 26

CHAPTER PAGE

V. THE FRAUD INVESTIGATION EXPERIMENT 27
(a) Background 27
(b) Experimental Design 29
(c) Expectations and Hypotheses 31
(d) Subjects 35
(e) Materials and Administration 37
(f) Manipulation Checks 40
(g) Results 42
Summary 54

VI. THE INTERNAL CONTROL RISK EVALUATION EXPERIMENT 56
(a) Background 56
(b) Experimental Design 57
(c) Expectations and Hypotheses 60
(d) Subjects 62
(e) Materials and Administration 63
(f) Manipulation Checks 64
(g) Results 65
Summary 70

VII. DISCUSSION 71
(a) Demographic Variables 71
(b) Conclusions 73
(c) Limitations and Suggestions for Future Research 77
Summary 80

- PAGE

REFERENCES 81

LIST OF APPENDICES 93

LIST OF TABLES

Table	Description	Page
I.	Components of Risky Choice Models	10
II.	Respondents by Position in the Firm	36
III.	Respondents by Level of Experience	36
IV.	"Even" versus "Odd" Responses	42
V.	Probability Estimates for C_{lsnt}	43
VI.	Responses to Probability Estimation Tasks for Groups C_{lst} and C_{lsnt}	44
VII.	Frequency of Resuscitations by Group	44

LIST OF FIGURES

Figure Page

1. The Fraud Investigation Experiment 30

2. Profiles of Final Mean Probability Estimates for C_{lsnt} 43

3. Mean Estimates of the Final Probability of Guilt of the Five Suspects 48

4. The Internal Control Risk Evaluation Experiment 59

5. Mean Estimates of The Final Probability of Internal Control Risk Categories 66

I. INTRODUCTION

Human decision making behaviour is a widely researched topic (for reviews see Einhorn and Hogarth [1981]; Pitz and Sachs [1984]; Rapoport and Wallsten [1972]; Slovic, Fischoff and Lichtenstein [1977]), and the question of how people make judgments under uncertainty occupies a prominent position in decision making research. One of the fundamental bases for this research has been the Probabilistic Judgment Paradigm. The importance attributed to this paradigm is illustrated by a quote from Fhaner: "A working knowledge of how people make probability judgments is essential for the application of any [decision] theories to human behavior [Fhaner 1977, p. 81]".

The exercise of (professional) judgment is a critical aspect of auditing.[1] Judgment under uncertainty, or risky decision making as it is known in the psychological literature, is an integral part of audit judgment.[2] In fact, the performance of an audit has been characterized as the practice of exercising judgment on the basis of incomplete and uncertain information [Mock and Vertinsky 1986, p. 1].

This study is an experimental investigation of the role of probability in auditors' decision making under uncertainty. (Familiarity with probabilistic concepts is identified by Ashton as one of the five necessary components of good audit judgment [R. Ashton 1983, p. 23].) Two experiments study fundamental concepts of axiomatic probability. The first is set in the context of Fraud Investigation.[3] The second experiment involves Internal Control Evaluation (see Felix and Niles [1988] for a review of judgment research into Internal Control Evaluation). The experiments examine the effect of (1) different response modes and (2) prior tutoring in axiomatic probability theory, on auditors' subjective probability assessments. Of particular interest is the effect of a visually-based graphical response mode. In two of the experimental conditions, the auditors are provided with such a response mode as a decision aid in order to allow them to generate their own overall probability mappings.

---

[1] The importance of judgment in auditing is evidenced by official pronouncements of both the U.S. and Canadian professional accounting bodies (i.e. SAS #3 [AICPA 1963, p. 34]; CICA HANDBOOK [CICA 1987/1968, sec 5000.04]).

[2] In the psychological decision making literature a risky decision is generally synonymous with a decision under uncertainty. Such decisions are usually operationalized as choices among several possible outcomes with known probabilities (e.g. gambles, lotteries). However, several authors (e.g. Edwards [1954]; Ellsburg [1961]; Knight [1921/1957]) have distinguished "uncertainty" as referring to situations where the probabilities of the alternative outcomes are unknown as well. This is more reflective of decision making in the auditing environment where the information is rarely, if ever, in explicit probabilistic form.

The graphical response mode is an important feature of these experiments. A widely-used standard reference on auditing states: "Graphic presentation of ... data facilitates ... evaluation." [Anderson 1984, p. 721]. The efficacy of graphical representation as a means of presenting information to subjects and facilitating their interpretation of such information has been examined (see Section Three, subsection (d) for a review of this research), but to the author's knowledge, this is the first time the question of the efficacy of graphical representation has ever been studied from the "other side" - the creation of one's own representation rather than the interpretation of someone else's.

One purpose of the experiments is to examine the presence or absence of probabilistic *thinking* in typical decision situations. The question of interest is whether the judge adheres to the basic axioms of probability in his or her probabilistic judgments. The experiments also examine some of the secondary characteristics of the judgments such as consensus and speed of decision making, which have been related in the literature to the efficiency and effectiveness of audit decision making, two components of professional practice which are very important in the current competitive market for audit services.[4]

---

[3] The topic of fraud investigation is a timely one. A recent recommendation by the Auditing Standards Board of the AICPA has expanded the auditor's responsibilities in the areas of detection and reporting of fraud [SAS #53, AICPA 1988], and in Canada the report of the MacDonald Commission of the CICA [CICA 1988] has just made similar recommendations.

Outline of the paper

Chapter Two describes the overall motivation for this research and its intended contribution – Chapters Three and Four provide background for this study. Chapter Three reviews the relevant literature from cognitive psychology pertaining to the role of probability in risk and risky decision making; Chapter Four provides an overview of the research in the field of audit judgment, especially research based on the probabilistic judgment paradigm. Chapter Five describes the Fraud Investigation experiment and Chapter Six describes the Internal Control Evaluation experiment. Chapter Seven contains the general discussion and conclusions.

---

[4] Since the experiments are intended to reflect the qualitative nature of the information with which the auditor deals, the quality of the judgments cannot be measured against any normative quantitative standards. Evaluation of the results is based instead on confidence in judgments and within-group agreement or consensus. Shanteau [1989] has argued that confidence is a necessary component of expertise and some authors (e.g. A. Ashton [1985]; Einhorn [1974]) have proposed that consensus can be a proxy for accuracy. It is suggested that any method which leads to greater agreement among experts can be viewed as a positive contribution. Besides the possibility of using consensus as a proxy for accuracy, consensus in audit decision making has been identified as a factor leading to increased efficiency and effectiveness in auditing [R. Ashton 1983, pp. 10–11].

II. MOTIVATION AND CONTRIBUTION

Audit judgment is an essential aspect of the professional practice of auditing (for reviews of the research in this area see R. Ashton [1982a, 1982b, 1983]; Birnberg and Shields [1989]; Burgstahler and Sundem [1989]; Davidson [1989]; Joyce and Libby [1982]; Libby [1981]; Solomon [1987]). This study examines audit judgment, focusing on auditors' use of probabilistic concepts in their decision making under uncertainty.

To this point in time, the normative Pascalian model of probability has been the basis for almost all of the research into decision making under uncertainty by auditors. The Pascalian paradigm is consistent with decision theoretic concepts of rational choice, and Bayes' Theorem, which is implicit in the Pascalian paradigm, is a logical consequence of conditional probabilities. The prescriptive literature (how should audit decisions be made) is also based on this normative Pascalian/Bayesian model of probabilistic decision making (e.g. Kinney [1975]; Larssen and Chesley [1986]; Leslie, Teitlebaum and Anderson [1980]; Sorenson [1969]). However, studies such as Joyce and Biddle [1981a, 1981b] have been interpreted as casting doubt on the descriptive validity of this approach, and studies such as Biggs, Messier and Hansen [1987] and Waller and Felix [1984] have suggested that audit judgments are more aptly characterized as "production systems" or condition-action response pairs (originally due to Newell and Simon [1972]).[5] The purpose of this study is to obtain additional evidence on the question of the descriptive validity of the Pascalian/Bayesian framework as a conceptual underpinning of audit decision making under uncertainty.

---

[5] In weighing the merits of these two views of audit decision making, it is useful to consider also whether the judgments can be considered to be "routine" or "non-routine", a distinction made by Gibbins [1984]. It may be that in routine judgments the process may be descriptively modelled by a condition-action response pair ("if" → "then"); however in non-routine or first-time judgments, where there has been no opportunity to "condition" the response, the probabilistic judgment paradigm seems more compelling.

Previous auditing research into probabilistic judgment has generally provided subjects with evidence in the form of items of information with numerical probabilities attached, bearing on the overall probability of some event. The ability of the subjects to combine this additional numerical evidence and assess its impact on the probability of the event of interest is then measured against the benchmark of Bayes' Theorem. Taken together, the evidence from these studies indicates that the characteristics or dynamics of many decision situations may interfere with people's normative use of probabilistic concepts.[6] The present research addresses this question from a somewhat different perspective in an attempt to circumvent problems which may have confounded the conclusions of the previous auditing studies. Although the weight of evidence appears quite negative, it may be premature to form conclusions about auditors' use of probabilistic concepts in their decision making processes. The interpretation of the results of these previous studies is rendered problematic by two factors stemming from the fact that subjects were required to manipulate quantitative probabilistic information as part of the experiments.

One major weakness of many of these previous auditing studies is a consequence of the artificiality of the decision situation relative to that which auditors experience in their working environment. Information rarely, if ever, comes in the form of items with probabilities attached, much less numerical probabilities. An obvious problem is lack of external validity – the results of the experiments do not tell us anything about auditors' real decision making processes. It is of questionable relevance to ask if auditors can manipulate this kind of data in conformance with Bayes' Theorem – the situation does not arise in practice. Therefore one aim of this research is to investigate whether auditors can use "typical" audit evidence or information in a manner which is consistent with the concepts of the probability framework, rather than to evaluate how they perform at the assessment and revision of quantititive probability values.

---

[6] Cognitive psychologists A. Tversky and D. Kahneman [1974] initially identified a number of decision making heuristics which could account for the biases in the subjects' responses (see also Kahneman, Slovic and Tversky [1982] for additional examples), and the results in the auditing literature are consistent with these heuristics and biases (e.g. Johnson [1983]; Joyce and Biddle [1981a, 1981b]; Swieringa, Gibbins, Sweeney and Larson [1976]).

This artificiality may also be a serious threat to the internal validity of the experiments. These experiments purport to study audit decision making by asking auditors to make decisions in an "auditing setting". However, if that setting is perceived as irrelevant to what the auditors have ever experienced in their actual practice, the validity of their responses is open to question. Some of them may have felt that the whole exercise was a foolish waste of time. Their responses may reflect that opinion and be meaningless. In order to minimize negative effects due to the perceived artificiality of the decision situation, the experimental materials were developed and refined in an extensive iterative process in concert with numerous different practicing professional accountants, independent of the experimental subjects. The intent of this procedure was to produce a set of materials which were more representative of actual auditing situations requiring judgment.

The second weakness, also directly related to the use of quantitative probabilistic information, may be responsible for the poor performance. Presenting the auditors with quantitative information and requiring them to combine it in accord with Bayes' Theorem is a cognitively complex experimental task. The complex arithmetical processing may have obfuscated evidence of adherence to Pascalian/Bayesian principles in their information processing behavior. Imposing a framework of mathematical probability on the participants requires them to process the probabilistic information in a manner in which unfamiliarity with the arithmetical complexity may be confounded with apparent unfamiliarity with the concepts of probability.

This study consists of two experiments involving probabilistic decision making in an auditing setting. The intent is to translate the orientation of the literature on risk and risky decision making from cognitive psychology to an applied auditing setting. A significant difference which is relevant to this translation is that while empirical risky decision research has been formulated and tested in the context of monetary gambles, with explicit probabilities of winning or losing known amounts, such is not the case in practical audit decision making. If auditors' uncertain judgment behaviour is in conformance with the kind of processes envisioned by the theoretical models, the information they evaluate should in fact be used to estimate the values of the alternative outcomes, the probabilities involved, or both.

This research focuses on the probability component. The experiments ask the auditors to (1) make initial probability assessments, and (2) revise those assessments on the basis of new evidence. However, rather than provide the auditors with quantitative probabilistic evidence, this study presents them with evidence which is in a form more typical of that which is available in many decision making situations. The evidence can be characterized as "qualitative" and "indirect" in that while it consists of facts which are perceived as pertaining to the likelihood of the event of interest, it is not a (series of) direct statement(s) about the event of interest containing varying quantitative degrees of uncertainty. The auditor must integrate and evaluate the evidence in order to arrive at a judgment, similar to the decision making environment the auditor faces in his or her professional practice. In addition, the auditor-subjects in two of the experimental conditions are provided with a visually-based graphical response mode to facilitate expressing their probability judgments. This graphical response mode is intended to provide a setting where the task of probability judgment is not confounded by requiring complex arithmetic processing, either to interpret the significance of the evidence, or to express the results of the probability judgment.

Experienced auditors can certainly be considered to have substantive expertise, which concerns knowledge about the domain in which the assessments are made, but not necessarily normative expertise, which concerns the ability to assess probability in conformity with the probability calculus [Winkler and Murphy 1968]. This lack of normative expertise, typical of most experts in specific applied domains, has two implications for present purposes. First, it is important to examine how an auditor maps his or her substantive knowledge onto a probability assessment. A sustantive expert, such as an auditor, may have difficulty expressing his or her substantive knowledge into a probability assessment, even after some deliberation. Although previous research has approached the problem as one of trying to elicit these already existing probabilities, Winkler [1967a] and Phillips [1970] have suggested that a decision maker has no personal probability which is there "for the taking" and Shafer and Tversky [1985] have argued that probability assessment should be viewed as a constructive process. The visually-based graphical response represents an operationalization of this constructive approach.

Winkler and Phillips also argue that it is important to recognize the auditor's uncertainty about whether his or her probability assessment adequately represents his or her substantive knowledge.[7] These experiments contrast the visually-based graphical response mode, representing the constructive approach of Shafer and Tversky [1985], with a traditional response mode, using quantitative linear probability scales to elicit probability judgments. The common judgments required of all subjects allow a distinction to be made between the effect of the two types of processing in terms of both how the auditor does map his or her substantive knowledge into a probability assessment, and about how well the auditor feels the resulting probability assessments properly represent his or her substantive knowledge.

Summary

Since the general population, and auditors as well, seem to be surviving quite well in spite of their apparent lack of conformity with normative standards of rational decision making, it has been suggested that probabilistic information processing does not capture peoples' uncertain decision making processes. However, it is quite possible that this apparent lack of conformity is a function of the experimental design rather than the underlying information processing. The standard against which subjects' responses are measured, because of its requirement for complex arithmetical processing, may be introducing an artifactual bias into the interpretation of the results. Therefore, by presenting the auditor subjects with qualitative information more characteristic of that which they evaluate in their usual audit decision making, this research examines probability estimation and revision at a more basic level – it investigates to what degree subjects will adopt a probabilistic framework in decision making situations where there is no requirement for complex arithmetical processing.

---

[7] This is especially true since the evidence in these studies can be considered to contain "second-order uncertainty", "ambiguity" or "vagueness" (respectively: [Ellsberg 1961]; [Einhorn and Hogarth 1985]; [Budescu and Wallsten 1987]). That is to say, the information is not conclusive or "certain" (first-order uncertainty) and the degree of diagnosticity is itself a matter of judgment (second-order uncertainty).

An examination of the role of probabilities in auditors' decision making may be beneficial in two ways. It can provide an indication of the generality of the approach taken by the commonly accepted models of risky decision making, and it may provide insight into how auditors do utilize the information they acquire in order to make judgments under uncertainty. By gaining a better understanding of how auditors make such decisions it may be possible to suggest improvements to the process to improve such factors as consensus and speed of decision making, which have been related in the literature to the efficiency and effectiveness of audit decision making.

III. PROBABILITY IN DECISION MAKING UNDER UNCERTAINTY

(a) Introduction

This study is an investigation into auditors' use of probability in their judgments under uncertainty. To illustrate the central role occupied by probability in the study of decision making under uncertainty, this section presents an overview of the approaches which have been proposed in the cognitive psychology literature (where it is referred to as "risk" or "risky decision making"). The following table lists the characteristics of the three components – outcomes, probabilities, and the combination rule – which have been suggested by the various models of risky choice.

Outcomes	Probabilities	Combination Rule
transitivity	completeness	multiplicative
dominance	positivity	additive
independence	additivity	cumulative
sure-thing principle	complementarity	

Table I. Components of Risky Choice Models

The characteristics of the probability component of the models can be expressed mathematically as:

1. $P(S) = 1$ (completeness)
2. $0 \le P(A) \le 1$ (positivity)
3. $P(\bigcup_i A_i) = \sum_i P(A_i)$ (additivity)
4. $P(\bar{A}) = 1 - P(A)$ (complementarity)

where:

S = the sample space

A = an event in S

$\bar{A}$ = the complement of A

A_i, $i = 1, 2, \ldots$ = a sequence of disjoint events in S

The first three expressions are described in the literature as the "axioms of probability" (Edwards [1971]; Lawrence [1967]; Yeh [1973]; Zehna [1969]), while the fourth expression is usually given as a thereom derived from the axioms. However some authors (e.g. Krickleberg [1965]; Wadsworth and Bryan [1960]) include both complementarity and the opposite of completeness (i.e. $P(\emptyset) = 0$) as part of the first axiom as well. The term "coherence" is used to characterize probability which obeys all the axioms and theorems. (See also Fishburn [1986].)

One way of classifying the upwards of two dozen models in the literature [Emby 1985; Libby and Fishburn 1977] is according to the method they use for combining outcomes and probabilities. Below is a brief review of some of the more common models of risk and risky decision making, under the headings multiplicative probability, additive probability and cumulative probability.

(b) Multiplicative Probability

The best known "family" of models uses multiplicative probability and is based on the concept of expected value (EV) which Edwards [1954] attributes to the 17th century mathematicians Pascal and Cardoze. The EV model arose from their studies of gambling at the court of Louis XIV. The expected value of a particular choice is the product of the objective value of the outcome multiplied by its objective probability of occurrence. The alternative with the highest product value is the appropriate choice. The overall EV model is:

$$EV = \sum_i p_i v_i$$

where:

p_i = probability of occurence of the ith outcome

v_i = value of the ith outcome

In the ensuing several centuries, the expected value approach has been refined somewhat. Bernoulli [1738/1967] and von Neumann and Morgenstern [1947] suggested replacing the (objective) outcome component with a subjective function representing the utility of the outcome to the decision maker and Edwards [1955] argued that the notion of objective probability in the expectation models was also too restrictive. He suggested the subjective(ly) expected utility (SEU) model which replaced the objective probabilities with a function representing the individual's subjective probabilities. Subjective probability is intended to capture the individual's degree of belief in the likelihood of a particular outcome. Nevertheless, the multiplicative form of all these expectation models is the same and the individual's subjective probabilities are assumed to follow the axioms.[1]

An alternative multiplicative approach to modelling risk and risky choice can be characterized as "moment" models. This approach is represented by the work of such researchers as Coombs [1969], Coombs and Bowen [1971], Coombs and Pruitt [1960], Lichtenstein [1965] and Pollatsek and Tversky [1970]. These models adhere to the restrictions on probabilities from the basic expectation model, but attempt to improve its accuracy by including additional moments of the probability distribution such as variance, skewness and kurtosis. This approach is also prevalent in the finance literature dealing with risk - for example the mean-variance model [Markowitz 1952]; mean semi-variance [Markowitz 1959]; confidence limit [Baumol 1963] and the probability of ruin model [Agnew et al. 1969]. The many variants of the moment models were combined in one formulation by Fishburn [1977] which is expressed as:

$$\text{Risk} = \int_{-\infty}^{x} (t - x)^{\alpha} \, dF(x), \qquad \alpha \geq 0$$

---

[1] Savage [1954] developed a more rigorous version of the SEU model based on the axioms of preference behavior such as transitivity, dominance, etc., from Table 1.

where:

t = target or aspiration level of return

x = actual return

α = a measure of importance to the decision maker
of deviations from the target return

F(x) = the cumulative distribution function of X

$\int_{-\infty}^{x} (\cdot)$ = the area of probability "below" the target return

The most recent of the multiplicative models is Prospect Theory [Kahneman and Tversky 1979a]. The model is made up of two functions which are combined into an overall Value function. First, the utility function of the EU model is replaced by a function which measures the value of the outcome from a reference point of zero or the status quo, rather than in terms of terminal wealth. Secondly, and of more interest to the present research, the (subjective) probability function is replaced by a "decision weight" function ($\pi(p)$), intended to capture the "... desirablity of prospects and not merely the perceived likelihood of these events [Kahneman and Tversky 1979a, p. 280]." However, Prospect Theory's π function deviates only very slightly from normative probability, and it is clear from Kahneman and Tversky [1979b] that they strongly support the Pascalian concept of probability as the appropriate criterion.

(c) Additive Probability

A somewhat different approach to modelling risk and risky choice is represented by the "risk dimensions" model [Payne and Braunstein 1971; Slovic 1967; Slovic and Lichtenstein 1968a, 1968b]. This model is based on a linear, additive combination of outcomes and axiomatic probabilities rather than a multiplicative combination rule. The model may be stated as:

$$A(g) = \alpha + W_1P_W + W_2\$_W + W_3P_L + W_4\$_L$$ [9]

---

[9] A limitation of this model is that it does not have the generality of the others. Its formulation is based on a binary outcome gamble, although extension to multiple outcomes would be straightforward.

where:

$A(g)$ = attractiveness of the gamble

α = a constant reflecting the individual's propensity for gambling

W_i, $i = 1,\ldots,4$ = weights attaching to each dimension

P_W = probability of winning

P_L = probability of losing

$\$_W$ = amount to win

$\$_L$ = amount to lose

(d) Cumulative Probability

Recently Lopes [1984] has proposed the "distributional" model. This model uses the objective values of outcomes and probabilities, but is based on the cumulative distribution of the frequency of outcomes rather than the probability density function.[10] The values of the outcomes are combined with the probabilities multiplicatively and the individual products are then cumulated. The cumulated values can be portrayed graphically by means of a Lorenz curve, which takes the shape of an exponential growth curve. The degree to which the curve deviates from (is concave below) a 45° ray from the origin is a measure of the riskiness of a particular gamble.[11]

(e) Empirical Results

Empirical tests of the expectation models have tended to focus on the characteristics of the outcome component, which appear to be oft-violated in practice (e.g. Becker and Brownson [1964]; MacCrimmon [1968]; Tversky [1969]) and have not addressed the nature of the probability component. The models have been formulated and tested in the context of choices among gambles with explicitly known outcomes and probabilities which conform to the axioms of probability.

---

[10] While Lopes' approach incorporates the ideas of personologists such as J. Atkinson [1957] and McClelland [1961], the inclusion of "personal characteristics" of the decision maker does not anticipate departures from the basic probability axioms.

[11] The approach is very similar to that employed in economics. Welfare economists use this method to assess the equality or inequality of the distribution of a country's wealth across its population. The more concave the Lorenz curve (especially at low cumulative values) the more unequal the distribution of wealth (cf. A. Atkinson [1970]).

Empirical tests of the alternative models indicated that they performed well in the domains in which they were developed. The risk dimensions model described people's choices among binary gambles well and empirical tests of the distributional model [Lopes 1984, 1986; Schneider and Lopes 1986] using lotteries showed that it explained choices among several alternative outcomes. However, subjects' estimates and manipulations of probability were not examined in these experiments either. The probabilities and outcomes were given to the subjects as part of the experimental design.

Summary

Probability, however it is combined with outcomes, is a fundamental component of the models of risk and risky decision making in the cognitive psychology literature. (Appendix A summarizes the major models of risk and risky choice according to their use of probability.) However, the range of situations appropriate to the study of uncertain decision making encompasses more than the domain of choices among gambles. This study specifically addresses the role of probability in a domain which involves a different type of uncertain decision, where the judge must construct his or her own probabilities based on the qualitative information available in such situations.

IV. JUDGMENT IN AUDITING

(a) Introduction

The purpose of this section is threefold. Its primary purpose is to review previous research into auditor judgment based on the Probabilistic Judgment Paradigm. However, prior to the review, a brief historical background will be provided on the genesis of research into audit decision making. Because these experiments also incorporate simple decision aids, this section will finish with a brief review of the development and use of decision aids in acccounting and auditing judgment research.

(b) Background

By the early-to-mid 1970's the study of human judgment had attained a prominent place in the field of cognitive psychology (for comprehensive reviews of the research at this time see Slovic and Lichtenstein [1971]; Rapoport and Wallsten [1972]). There existed mature accepted scientific paradigms, such as the policy-capturing paradigm, based on Brunswik's Lens Model [Brunswik 1952, 1956], the probabilistic judgment paradigm, based on the concept of expected utility [Bernoulli 1738/1967; von Neumann and Morgenstern 1947] and the principles of decision theory [Raiffa 1968], which could be adapted to the applied field of audit decision making.

The paradigm first employed in audit judgment studies was the the policy capturing paradigm. A major objective of this type of study is to identify the manner in which auditors combine and weight various cues (items of information) in order to make their judgments. Another objective of policy capturing research is to enable the researcher to construct mathematical models which will reproduce (and often improve on) the auditor's judgments. The general methodology is for the researcher to present the auditor/judge with a set of cases in which the facts (cues) differ systematically from case to case.[12] Analysis of variance (ANOVA) or multiple regression analysis (MRA) can then be used to statistically analyze the results.

The first substantive auditing issue studied using this paradigm was internal control judgment. The ANOVA approach was used by R. Ashton [1974a, 1974b] to study internal control judgments. By analyzing how auditors' judgments differed from case to case as the values of the cues changed (i.e. the presence or absence of specific controls), he was able to infer how influential specific aspects of the client's internal control system were in the judgments. This approach to the study of internal control was subsequently replicated or extended by numerous other researchers – e.g. Abdel-khalik, Snowball and Wragge [1983]; Ashton and Brown [1980]; Ashton and Kramer [1980]; Gaumnitz, Nunamaker, Surdick and Thomas [1982]; Hamilton and Wright [1982]; Joyce [1976]; Mock and Turner [1979]; Reckers and Taylor [1979]; Weber [1978]; Wright [1982].

Other areas which have been investigated using the ANOVA methodology are internal audit evaluation [Gibbs and Schroeder 1979], materiality judgments [Moriarity and Barron 1976, 1979; Hofstedt and Hughes 1977; Firth 1979], and audit report messages [Libby 1979a, 1979b].

MRA is another statistical technique based on the general linear model. The dependent variable for the regression is the judgment about the quality of the overall internal control system, and the independent variables are the values of the specific controls. It permits construction of mathematical models which will reproduce the judgments of subjects to a very high degree of accuracy. (In fact, the predictions of the MRA model may be more valid than those of the judges whose judgments were used to construct/generate the model in the first place – see for instance Dawes and Corrigan [1974]). The MRA model is not as appropriate as a statistical technique for the purpose of explanation of judges' decision making strategies; rather it is most suited to the creation of predictive models of decision making.[13] A major application of this technique has been to build models of bankruptcy prediction based on the financial ratios of companies [Altman 1968, 1982, 1983].

---

[12] Stictly speaking it is the values of the cues, rather than the cues themselves which differ from case to case.

[13] For instance, the measurement and interpretation of the relative importance of various cues in this model is somewhat problematic – see R. Ashton [1983, pp. 21-3]

A major drawback of the policy capturing paradigm is that it does not explicitly consider the issue of uncertainty, a critical feature of much audit decision making. Wright [1982] argued that while for certain types of judgment research, the simpler lens model approach is satisfactory, for investigations dealing explicitly with the effects of uncertainty, the subjective probability model, based on the SEU paradigm, is superior. The following section reviews the research that has been done in auditing judgment based on this approach.

(c) Probabilistic Judgment

The other major paradigm that has guided research in the area of audit judgment is based on the concept of Subjective Expected Utility [Edwards 1955; Savage 1954]. SEU is the foundation for a theory of rational choice called decision theory [Raiffa 1968] which provides a framework for the study of decision making under uncertainty. Within this approach the major focus of research has been on developing and describing techniques that will provide a good measure of judges' subjective beliefs (i.e. subjective or personal probabilities). The research into probabilities can be further broken down into (i) probability elicitation, (ii) belief revision, and (iii) probability manipulation.

(i) Probability elicitation

This stream of research is concerned with investigating assessments of probabilities by judges to reflect their own subjective beliefs. It is based on the assumption "... that auditors must, explicitly or implicitly, assess and use probabilities..." when making many different types of audit judgments [R. Ashton 1983, p. 23].

The objective of probability elicitation research is to have judges estimate a probability distribution for an event of interest which corresponds to their belief in its likelihood of

---

[13](cont'd) for a discussion.

occurrence.[14] In auditing, one stream of probability elicitation research has been concerned with comparing different methods of eliciting assessments of probabilities by practicing auditors to reflect their beliefs about error rates in accounting populations ("attribute" contexts). Felix [1976] compared two elicitation techniques – "quartiles" versus "equivalent prior sample (eps)" – for assessing the probability of errors in a hypothetical order receiving/shipping/billing system, and Crosby [1980, 1981] added an additional step by having the auditors use their assessed probabilities to estimate appropriate sample sizes. He first compared assessed distributions using "eps" versus direct estimation of the cumulative distribution function in a sales/accounts receivable/collections cycle. After the subjects completed the probability assessment task they then made judgments as to the appropriate sample size for subsequent audit testing. The recommended sample sizes using the two different techniques were widely different, and compared to the normative Bayesian standard, both of the elicitation methods led to excessively large sample sizes. Kinney and Uecker [1982] compared the direct fractile elicitation method versus a risk or probability assessment method to estimate population error rates. As a result of their study, Kinney and Uecker suggested that the risk assessment method might be preferable in the context of internal control evaluation. Chesley [1976, 1977, 1978, 1979] and Wright [1980] also studied the use of different probability elicitation techniques (although their experiments used abstract tasks and student subjects as opposed to the above-mentioned work).

Researchers such as Solomon [1982]; Solomon, Krogstad, Romney and Tomassini [1982]; Wright [1989] have used various elicitation techniques to obtain probability assessments for the values of account balances ("variables" contexts). Tomassini, Solomon, Romney and Krogstad [1982] used the fractile method in the estimation of accounts receivable and inventory balances. Abdolmohammadi [1986] summarized this stream of research. His review indicated that although the results of this research are in some ways inconclusive and conflicting, a point made earlier by Ashton [1983] and Libby [1982], generally speaking the results were that different elicitation methods produced quite different assessments and that the assessments, both within and between elicitation techniques, were often quite inconsistent.

---

[14] The roots of this research in auditing appear in the cognitive psychology literature in the latter 1960's and early 1970's – e.g. Du Charme and Donnell [1973]; Schaefer and Borcherding [1973]; Stael von Holstein [1970, 1972]; Winkler [1967a, 1967b, 1971].

A subset of the probability elicitation research has been concerned with the accuracy of the elicited probability distributions. Accuracy has been operationalized by the standards of calibration – the evaluation of the subject's probability assessments by comparing them to the actual frequency with which the event occurs – and extremeness – the tightness or peakedness of the subjectively assessed probability distribution.[15] Abdolmohammadi and Berger [1986] concluded that an elicitation technique based on the probability density function was superior in terms of accuracy, but the standard against which the elicited probabilities of their subjects was measured was the result of a 600 sample simulation rather than actual external frequency data. The previously cited studies by Solomon [1982] and Tomassini et al. [1982] used both these measures to assess the goodness of the auditors' elicited probabilities. They had practicing auditors provide assessments for the audit values of six different accounts (four accounts receivable, two inventory) taken from Neter and Loebbeke [1975]. Both studies found that the auditors' assessments tended to be miscalibrated in the direction of underconfidence (seeing events as being less likely than they really are) and extreme (too "tight"). Libby [1981] reasoned further that in the context of compliance testing, overconfidence could lead to under–auditing and impair audit effectiveness and underconfidence could lead to over–auditing and impair audit efficiency.

An experiment by Beck, Solomon and Tomassini [1985] concluded that the effects of miscalibration in the assessment of the probability distribution of account balances on the resulting level of audit risk was complex and depended on the nature and severity of any extremeness deficiencies. In other words, overconfidence could lead to impairment of efficiency or effectiveness, as could underconfidence. In summary, the bulk of evidence consistently indicated that the auditor–subjects appeared to have little facility with converting and processing information in quantitative probabilistic form.

---

[15] This research again has its roots in the cognitive psychology literature – for a review see Lichtenstein and Phillips [1977]. It has not been as widely replicated in auditing, perhaps because of the difficulty of obtaining objective data from "out there" with which to compare auditors' subjective probability assessments.

(ii) Probability revision

Another sub–topic closely related to probability assessment is concerned with judges' revisions of probability estimates. Very often an initial judgment is revised on the basis of additional evidence. Such probability determinations are a necessary precursor to the application of the models of decision making discussed in Chapter Three. Normative decision making in a dynamic environment, or multi–stage decision making (also known as cascaded inference – see Schum and Martin [1982]) is prescribed by Bayes' Theorem, which is expressed in discrete form as:

$$P(H_i|D) = P(D|H_i)P(H_i)/P(D)$$

where:

H_i – the hypothesis under consideration

D – the new information

Bayesian revision is central to many areas of accounting research, such as principal–agent modelling [Baiman 1982], information economics analyses [Demski 1980], and audit evidence evaluation [Scott 1984].[16]

Probability revision by individuals has been examined in a number of accounting–related settings. Dickhaut [1973] studied the probabilistic estimation of business profitability, Kennedy [1975] examined probability revision in the context of predictions of bankruptcy versus non–bankruptcy, and Wright [1979] studied probabilistic revisions of β_i, the systematic risk of a security. Their findings were consistent with those of the psychological literature. When the subjects' performance was evaluated by comparison to the objective results obtained using Bayes' Theorem, the principal finding was conservatism. Subjects typically revised their estimates in the appropriate direction, but the extent of the revision was insufficient relative to the prescription of Bayes' Theorem [Edwards 1968].

---

[16] Bayesian revision was introduced in the psychological literature by Edwards, Lindman and Savage [1963], and some of the earliest empirical work was done by Edwards, Lindman and Phillips [1965]. The field is reviewed extensively by Slovic and Lichtenstein [1971].

In a recent auditing-related study, Abdolmohammadi [1986] examined auditors' revision of sample sizes for substantive testing based on the results of compliance testing. He compared auditors' actual sample size decisions with those that would have resulted from the application of Bayes' Theorem. The predominant result was again conservatism, the auditors did not reduce sample sizes as much as was warranted by the compliance testing results.

(iii) Probability manipulation

The third measure of "statistical rationality" [R. Ashton 1983, pp. 9-10] is coherence (the first two are calibration and extremeness), which refers to the degree to which subjective probabilities obey the laws or axioms of probability (see also above - Chapter Two, page 10). With the exception of Chesley [1984], the coherence axiom has not received much attention in the accounting and auditing literature. Tests of subjects' abilities to assess and use probabilities have been done against the benchmark of mathematical standards (Bayes' Theorem). The experiments have been designed in such a way as to constrain subjects to operate within the bounds of axiomatic probability by using experimental settings in which probability values are given to subjects as part of the experimental design.

The role of probability in "natural" uncertain decision making processes has not been investigated. The question of whether or not subjects use probabilistic information normatively when it is presented to them in numerical form is somewhat different from the more basic question of how they use information to assess probabilities. Since most information in typical uncertain decision making situations is not in quantitative form in the first place, it is appropriate to step back and examine how (or if) the concepts of probability are applied to the processing of this information. To date the topic of direct probability manipulation has not been addressed in the auditing judgment literature. This subsection reviews the research in the cognitive psychology literature as an empirical foundation for the present study.

Among the first tests of the axioms of probability were those by Teigen [1974a, 1974b]. His subjects consistently violated the axiom that the probabilities of a set of exhaustive and mutually exclusive events should sum to 1.0 or 100%. They produced probabilities which summed on average to approximately 2.5, with the degree of overestimation tending to increase as the number of alternatives increased. Another series of experiments [Teigen 1983] solidly supported this finding. As the number of alternatives increased, the sum of the subjects' initial probability estimates increased also, from about 1.3 with two alternatives, to about 2.6 with 10 alternatives. He termed this a non-distributional concept of probability and added the caveat that when subjects are instructed to make their estimates sum to 1.0, one should be very cautious about interpreting the results as reflecting subjective probabilities.

A paper by Robinson and Hastie [1985] extended these ideas to look at the coherence of people's probability revisions following the receipt of new evidence. For subjects untutored in probability theory, initial estimates tended to sum to approximately 2.0 (with eight alternatives), and increased or decreased in total as new items of evidence were introduced. In other words, the adjustment process for the probability of one hypothesis was insensitive to evidence regarding competing hypotheses, consistent with the non-distributional concept of probability in Teigen [1983].

The behaviour of a group of subjects in Robinson and Hastie [1985] who received tutoring in probability theory prior to the experiment did conform to the completeness and complementarity constraints. However, such conformity did not appear to come easily. The researchers noted other strategies which the subjects employed, such as premature elimination of alternatives, in order to keep the task manageable. As a consequence, the tutored subjects engaged in frequent "resuscitations" - an alternative which had been assigned a zero probability on the basis of the evidence to date was reassigned some positive probability following the receipt of additional new evidence.[17]

---

[17] This directly contradicts acceptable patterns of Bayesian revision. Once the probability of an alternative is reduced to zero it is a permanent state. Consideration of the formula which describes Bayesian revision, $P(H_i|D) = P(D|H_i)P(H_i)/P(D)$, shows clearly that once $P(H_i) = 0$, the numerator, and consequently the value of the entire expression will subsequently always be equal to zero.

(iv) Summary of probability studies

It is clear that although probability estimation and manipulation are central to decisions under uncertainty as modelled normatively, the evidence indicates that neither the general population nor auditors are very good at it. The probability elicitation studies show that subjects were not comfortable with a quantitative probabilistic mode of information processing, and this is reinforced by the fact that where external referents were available, their performance as measured by calibration was poor. The tendency of individuals to be conservative in their probability revisions, originally demonstrated by Edwards and his associates, and subsequently also revealed in an auditing context [Abdolmohammadi 1986; Joyce 1976] also indicates a lack of facility with quantitative numerical manipulation. Although other authors have offered other explanations of the experiments which demonstrated conservatism (e.g. Beach [1968]), or criticized the experimental paradigm itself [Peterson 1968], certainly one possible explanation is that the concept of probabilistic information processing is not consistent with people's decision making processes. However, the interpretation of many of the studies into subjects' use of probability is rendered problemmatic by the fact that the subjects were required to undertake complex arithmetical processing as part of the experimental design. Unfamiliarity with the arithmetic may have been mistakenly interpreted as unfamiliarity with patterns of human information processing which in fact were consistent with the Pascalian/Bayesian framework. This study attempts to circumvent that problem. It examines subjects' use of probability at a more basic level, without requiring them to perform the complex arithmetical processing such as would be involved in quantitative Bayesian revision.

(d) Graphical Decision Aids

One method which has been suggested as a means of helping to improve human judgment is to provide the judge with some form of decision aid. Because the present study makes use of simple decision aids to assist the auditors in making their probability judgments, the following section reviews some of the research which has been done on this topic, especially that using graphical representations of quantitative data, which is the spirit of this research.

One of the earliest examples of the use of graphical decision techniques to represent quantitative empirical data appears to have been by Playfair [1786]. He is credited with inventing the line graph, the bar chart and histogram, and the circle graph or pie chart, among others. The question of the relative efficacy of these various forms of graphical representation was not really addressed for some years, although there was a general feeling that graphical data representation illuminated relationships and/or trends in the data that tabular presentation often obscured. [18]

In the early 20th century, research was done on the usefulness of different forms of graphical presentation by Croxton and Stein [1932], Croxton and Stryker [1927], and Eells [1926]. These researchers concluded that the pie chart form of representation was at least as good as, and in fact in almost all cases superior to, any alternative graphical method of data presentation.[19]

Research in accounting and auditing judgment using graphical decision aids has been limited, but one type of graphic representation to have been adapted from the psychological literature to accounting research is from Chernoff [1973].[20] "Chernoff Faces" as they have become known, were used by Moriarity [1979]. He found that subjects experienced in financial analysis who were required to make judgments of bankruptcy on the basis of financial data, were more accurate when the information was presented graphically than when it was presented in the traditional quantitative format. His results were supported by Stock and Watson [1984] using accounting faculty and PhD students in an experiment involving bond rating. They found that the subjects receiving the graphical presentation were more accurate than those receiving the information in quantitative form and that the graphical presentation alone was as good as the graphical presentation plus the quantitative data.

---

[18] A quotation from two statisticians of the late 19th century puts it very succinctly: "Getting information from a table is like extracting sunlight from a cucumber [Farquhar and Farquhar 1891]."

[19] Research into risky decision making in the cognitive psychology literature has also made use of the pie chart form of presentation of probabilistic information (see for instance Payne and Braunstein [1971]; Slovic and Lichtenstein [1968a, 1968b]; Budescu and Wallsten [1988]; Wallsten et al. [1986]).

[20] Research in cognitive psychology using graphical decision aids is reviewed by Wainer and Thissen [1981].

Summary

Judgment under uncertainty is a fundamental aspect of professional auditing judgment. The extent to which the probabilistic judgment paradigm represents auditors' information processing behaviour in judgments under uncertainty is the focus of this research. The intent is to provide the judge with a simple non-quantitative decision aid which will assist him or her in constructing coherent probability assessments. It is hoped that such a decision aid will be beneficial in terms of promoting a descriptive model of uncertain decision making which follows the concepts of the normative Pascalian/Bayesian model of probabilistic judgment, while at the same time enhancing secondary characteristics of the judgment.

V. THE FRAUD INVESTIGATION EXPERIMENT

(a) Background

The first experiment draws upon and extends the methodology of Robinson and Hastie [1985] (hereinafter RH). Their study examined university student-subjects' likelihood estimation and probabilistic belief revision processes. The subjects were first asked to estimate their subjective likelihoods of guilt for five suspects in a brief murder mystery. They were then provided with additional items of evidence, one at a time, regarding the five suspects, and revised their subjective likelihoods each time on the basis of each of those additional pieces of evidence. The evidence was of the type that would be characterized as "qualitative" and "indirect" according to the definitions in Section Two.[21] RH were not examining subjects' ability to combine numerical probabilistic evidence according to normative prescriptions, rather they were studying subjects' conformance or non-conformance with the probability axioms of completeness and complementarity. The subjects' responses were compared to the patterns prescribed by the Pascalian/Bayesian probability framework.

Their subjects were divided into two groups, one "naive" group who had no formal training in probability theory, and one group who received tutoring in probability theory (focusing on the axioms) prior to the experiment. The experiment contrasted the propensities of the naive and tutored groups to adhere to the normative framework in their probability estimates and revisions.

The naive subjects violated the axioms of completeness and complementarity. The tutored group represented an attempt to "cure" the subjects of their violations of the axioms by tutoring them in axiomatic probability theory prior to the estimation and revision. The tutoring was successful in that it eliminated the violations, but it resulted in the subjects engaging in frequent resuscitations, which directly contradict Bayesian patterns of belief revision. However, the subjects

---

[21] In legal terms, this evidence would probably be defined as "circumstantial" - it related to motive, means, and opportunity, but was not a direct statement of the probability of guilt of any suspect.

were required to express their probabilities of guilt for the five suspects on a set of five linear probability scales each time, and it is the contention of this study that the resuscitations were more apparent than real. The resuscitations witnessed by RH were, rather than evidence of non-Bayesian processing on the part of the tutored subjects, due to the fact that this experiment also required considerable intermediate arithmetical processing. The scales were anchored by "0.0 – no chance" and "1.0 – sure thing", and were divided into 10 equal segments. While the subjects did not have to combine numerical probability statements to be evaluated against the prescriptions of Bayes' Theorem (the new information was qualitative rather than quantitative), they were required to allocate their revised probabilities among the five suspects so that the sum of the five scales was equal to 1.0 each time. Thus they would have to keep track of how much probability they had "used up" and how much they "had left" in using the response scales each time. The order of the suspects was also randomized for each repetition, which would further increase the need for "mental bookkeeping". This provides a plausible alternative explanation for the resuscitations. Tutored subjects who eliminated one or more suspects very early on (i.e. $P(A_i) = 0$) may have done so to facilitate completeness and complementarity in their subsequent estimations by reducing the number of alternatives they had to consider at each iteration, rather than because they were certain that the suspect was not guilty. The ensuing resuscitations (i.e. $P(A_i) > 0$) could be a result of these premature eliminations.[22]

In the Fraud Investigation experiment the auditors were asked to estimate and revise their subjective likelihoods of guilt of five suspects in a fraud investigation. There were nine iterations – the initial situation and eight additional pieces of evidence.

It should be stressed that the nature of the evidence was critical to both the internal and external validity of this experiment. It was important that the evidence be representative of the kind of information upon which the auditor must formulate his or her decisions in this type of investigation. Such evidence is generally qualitative and indirect rather than quantitative and direct

---

[22] This explanation is even more plausible if one considers that the experiment contained five suspects and 13 pieces of evidence – for a total of 65 individual probability estimations.

– it must be evaluated and integrated by the auditor to infer the probability of guilt of the suspects rather than being manipulated according to some algebraic formula. It is the contention of this study that the negative conclusions of much of the previous research into probabilistic decision making and Bayesian revision were confounded by the presence of an intervening variable – the requirement for subjects to engage in complex arithmetical calculations as part of the experimental task. Although previous experiments have demonstrated that subjects were unable to conform to the mathematics of Bayesian revision, the Pascalian/Bayesian probability framework upon which the normative audit decision literature is based may still be valid as the conceptual underpinning of a model of auditors' decision making under uncertainty. (See for instance R. Ashton [1983, p. 2]; Leslie, Teitlebaum and Anderson [1980, p. 303].)

(b) Experimental Design

This experiment utilized two experimental treatments, each having two levels.The first factor was the response mode (five linear scales per iteration – one scale per suspect (Appendix D.3) – versus a single circle graph per iteration (Appendix D.4)). The second factor was the level of prior training. One level was prior instruction in axiomatic probability (Appendices B.1 and B.3), the other level was no prior instruction. The resulting experimental design was a completely crossed 2 X 2 factorial with four different experimental conditions:

C_{lst} – linear response scale with tutoring
C_{lsnt} – linear response scale with no tutoring
C_{pct} – pie chart response mode with tutoring.
C_{pcnt} – pie chart response mode with no tutoring

The Fraud Investigation experiment, which is of the form described as the post-test only design [Campbell and Stanley 1966, pp. 25-26] is presented schematically below in figure one:

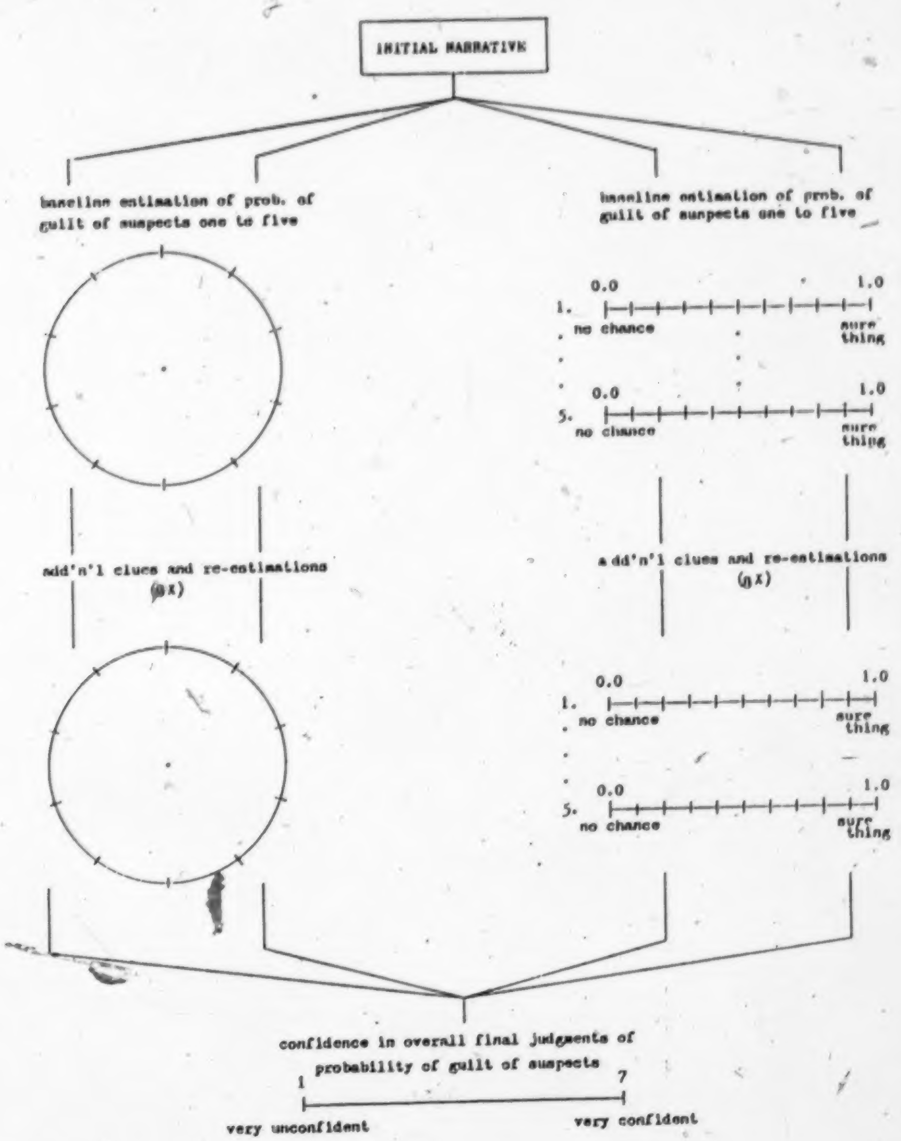

INITIAL NARRATIVE
baseline estimation of prob. of guilt of suspects one to five
baseline estimation of prob. of guilt of suspects one to five
0.0
1.0
1.
no chance
sure thing
5.
add'n'l clues and re-estimations (n x)
confidence in overall final judgments of probability of guilt of suspects
1
7
very unconfident
very confident

Figure 1. The Fraud Investigation Experiment

Conditions C_{lst} and C_{lsnt} allowed direct tests of the completeness and complementarity probability axioms (as in RH) in an auditing setting. The objective was to investigate whether these subjects were prone to the same errors and violations as student subjects. It was expected that in general the results would be similar to those of the previous research.

This experiment extended RH by adding the additional experimental treatment or condition of a pie chart response mode for subjects' responses (conditions C_{pct} and C_{pcnt}). This automatically ensured that their responses would satisfy completeness and complementarity.

The purpose of this additional factor was to facilitate a comparison of the effect of different modes of processing on characteristics of the judgments such as consensus and confidence. Subjects using the pie chart response mode would not be required to engage in arithmetical processing in either the input stage or the output stage of their judgments. Thus the pie chart response mode represented an attempt to completely eliminate the presence of the intervening variable of arithmetic processing.

(c) Expectations and hypotheses

It was the general expectation of this experiment that removing the need for intermediate arithmetic processing would allow the subjects to concentrate more effectively on the evidence itself and would result in an improvement in characteristics such as confidence and consensus. With respect to confidence, subjects with different degrees of facility and comfort in expressing their judgments could have different levels of confidence in those judgments. With respect to consensus, the fact that some of the subjects were required to follow the arithmetic "rules" might interfere with their judgment processes and might introduce some extra variation in their judgments, over and above the normal degree of within-group variation.

As discussed above, one of the questions towards which these experiments are directed, is whether or not auditors, in an audit decision making situation, follow the same information processing patterns as shown by the subjects of RH.[23] Accordingly, the first two hypotheses relate to the previous results in this area – e.g. Robinson and Hastie [1985]; Teigen [1983]. They concern the prevalence of violations of the axioms of completeness and complementarity and the frequency of resuscitations. In general the expectations are that the results will corroborate those of prior studies. These expectations may be formally stated as:

> H_1: The subjects in C_{lst} will comply with completeness and complementarity; the subjects in C_{lsnt} will not.[24] [25]

(Since the response mode does not lend itself to precise estimates, some deviation from 1.0 is to be expected. However, if subjects' estimates are significantly greater than 1.0, that can be interpreted as a violation of the completeness and complementarity axioms.)[26]

> H_2: The subjects in C_{lst} will exhibit the most resuscitations, followed by the subjects in C_{lsnt}. The subjects in C_{pct} and C_{pcnt} will not exhibit resuscitations.

---

[23] There is limited evidence in the literature that auditors are not as prone to heuristics and biases in some areas of audit decision making as the subjects in cognitve psychology research (e.g. A. Ashton and R. Ashton [1988]). However most of the research to date on the conformance or non-conformance of auditors to normative standards has found that their decision processes are similar to those of non-specialized subjects (e.g. Gibbins [1977]; Swieringa et al. [1976]).

[24] Completeness refers to the initial assessments of the probability of guilt of the five suspects and Complementarity in this context refers to the requirement for compensatory changes in the assessments of the probability of guilt of the five suspects with the receipt of new evidence.

[25] Recall that the response mode for the other two groups guaranteed compliance with the axioms.

[26] Shafer [1976] and Shafer and Srivasta [1987] have proposed "belief functions", a form of probability which does not require the decision maker to maintain $P(S) = 1$ (or equivalently $\Sigma P(A_i) = 1$). However, both the belief function approach and the inductive probability of Cohen [1977] contemplate circumstances where it is acceptable to have $\Sigma P(A_i) < 1$, but do not deal with the situation where $\Sigma P(A_i) > 1$, which has been found to be the usual form of violation [Teigen 1974a, 1974b].

If removing the need for intermediate processing is in fact removing a potential source of interference, and if subjects' intuitive decision making processes do follow the patterns of Pascalian/Bayesian probability theory, it could be expected that subjects using the pie chart response mode will not engage in resuscitations. The subjects in C_{lsnt}, who use the same linear scales as C_{lst} but who are not explicitly constrained by the experimental instructions, are expected to be "in the middle".

The next two hypotheses relate to the expectation that the use of this visually-based decision aid will facilitate these subjects' (C_{pct} and C_{pcnt}) ability to use this information to make probabilistic judgments.

Torgerson [1958] discussed the relative merits of different modes of quantitative data presentation within the framework of different scaling techniques. He suggested that the difficulty or ease of use of different forms of data presentation could be visualized in terms of scale complexity, which he described as a function of the number of cognitive scale elements (e.g. units, origin, distance) which must be processed and integrated. By extension, one could expect that a single circle graph, as opposed to a series of linear scales, would be superior in allowing auditors to express their probability judgments.

A list of criteria proposed by Gnanadesikan [1980] for judging the relative merits of various alternative graphical modes, such as descriptive capacity, data orientation, potential for internal comparisons, aid in focusing attention, and adaptability to large volumes of data, also supports this expectation and Bowman [1968] suggests that the benefits of this approach come from its ability to show "... separate segments in relation to a visually unified whole... (p. 173)". In a more related field, this expectation is also consistent with the findings in the Management Information Systems literature (e.g. Ives [1982]; Remus [1987]). Chernoff [1978] suggests that decision aids which have a strong visual component are beneficial in portraying "... a more complete and better-balanced understanding... (p. 4)." In the context of the judgments required in these experiments, the use of a visually-based decision aid which stresses the simultaneous rather than sequential nature of the probability estimations - one in which the relationship between the alternatives is made explicit – could be expected to be superior.

It is expected that, in general, the results will reflect this enhanced ability of the subjects to interpret the implications of the evidence with respect to the probability of guilt or innocence of the suspects. If in fact the responses are consistent with the expectations which may be inferred from the related literature, the responses of the subjects in the four treatment conditions will follow a similar pattern in the two hypothesis tests. In other words, it is expected that the complex arithmetical calculations will "get in the way" of the assessments of the subjects using the linear scales and receiving the tutoring, and that they will exhibit the least amount of agreement amongst themselves in H_3 and the lowest mean rating in H_4. The subjects in the pie chart conditions will exhibit the greatest agreement/highest rating. (It may well be that for the subjects using the pie–chart response format, tutoring in probability theory is not necessary and there will not be any significant difference between the performance of the tutored subjects and the performance of the untutored subjects.) If the initial tests indicate a significant difference between conditions, *post hoc* comparisons such as Newman–Keuls [Myers 1979, pp. 296–98] can be used to examine the order of the responses (e.g. $C_{pcnt} \geq C_{pct} > C_{lsnt} > C_{lst}$). These expectations may be formally stated as:

> H_3: The amount of within–group variance in the assessments of the probabilities of guilt of the five suspects will be least in the pie chart condition and greatest in the linear scale tutored condition.[27]
>
> H_4: The confidence ratings of the subjects will be highest in the pie chart conditions and lowest in the linear scale tutored condition.[28]

The fifth hypothesis tests the expectation that the differing response modes and restrictions on the subjects' judgment processes would result in differing perceptions of the difficulty of the probability estimation task. The same considerations apply to order effects with respect to H_5 as with respect to the previous two hypotheses. This expectation may be stated as:

---

[27] See Section I, footnote 4 regarding measurement criteria.

[28] Ibid.

H_5: The subjects in the pie chart conditions will perceive the probability estimation task to be less difficult than the linear scale conditions.

(d) Subjects

The subjects for the experiments were 140 practicing auditors from 20 CA firms in the Vancouver area. The initial request for participation was made by letter to the personnel partner. The letter explained the general nature of the study, the approximate time required to complete it, and the "target" population (auditors with at least one year's field experience). This was followed up by telephone approximately one month later. If the personnel partner agreed to the firm's participation (and all but two very small firms did), a tentative time was arranged for the experiment. The personnel partner contacted other members of the firm to enlist their co-operation. Several days before the arranged time the personnel partner was again contacted by phone to confirm the arrangements and so that sufficient experimental materials could be brought for the number of participants in the session. In a few cases, pressing business matters made it necessary to reschedule the session, but this was not a frequent occurrence.

The subjects completed the experiment in groups of three to 20, depending on the size of the firm. Some sessions were scheduled immediately following in-house training sessions or staff meetings, some were scheduled during the work-day and some were scheduled at the end of the regular business day. I administered all 140 cases personally, providing the subjects with the same pre-experimental instructions, tutoring, etc. (Appendices B.1 - B.4, C)

The purpose of the experiments was explained to the subjects as "to see how people make decisions based on incomplete information". There was no subterfuge or misdirection in the presentation of the experiments.

At the end of the experimental session the subjects were asked to furnish certain demographic background information, such as length of experience, position in the firm, educational background, whether they had ever been involved in a fraud investigation, etc.

(Appendix F). This information was obtained in order to ensure the equivalence of the four experimental groups and also to investigate other possible relationships in the responses. The subjects were assured of the anonymity of their responses. The participants were well experienced, with a strong representation from the supervisor/manager level (46%) as well as from the partnership (31%) level. Tables II. and III. below present descriptive statistics on the participants:

	Condition			
	C_{pct}	C_{pcnt}	C_{lst}	C_{lsnt}
Position				
Staff	7	8	9	8
Sr./Sup.	18	15	16	15
Mgr./Ptr.	10	12	10	12

Table II. Respondents by Position in the Firm

	Condition			
	C_{pct}	C_{pcnt}	C_{lst}	C_{lsnt}
Experience				
≤ 2yr	10	13	10	10
$> 2 < 6$	19	14	17	13
≥ 6	6	8	8	12

Table III. Respondents by Level of Experience

The proportions of respondents were equivalent across the four experimental conditions by either grouping (Table II.– $\chi^2 = .999$, $p \simeq .98$; Table III.– $\chi^2 = 4.308$, $p \simeq .66$).

(e) Materials and Administration

The subjects were randomly divided into four treatment conditions, as described above. The subjects in C_{lst} received prior training in probability theory consisting of training in the permissible range of probabilities (0.0 to 1.0) and in the additivity of probabilities for mutually exclusive and exhaustive events (Appendix B.1). As part of the training these subjects worked through an exercise with the experimenter similar to the experiment but on a different topic (a murder mystery - see Appendix C) involving initial probability estimation and one re-estimation on the basis of additional evidence. The training exercise introduced the response scale to be used in the experiment (a 10 cm. scale running from "0.0 - no chance" to "1.0 - sure thing", marked off in 10 intervals of one cm.) and was used to illustrate the initial estimation of probabilities, and one re-estimation, according to the axioms of completeness and complementarity.

The subjects in C_{lsnt} received only a basic introduction to the concepts of probability, with no reference to the completeness and complementarity axioms (Appendix B.2). However, they did work through the same exercise as C_{lst} (Appendix C), involving estimation and re-estimation. The purpose of the exercise was simply to familiarize them with the format of the experiment. The exercise introduced the response scale (the same as for C_{lst}) but did not describe how it "should" be used, to avoid influencing the subsequent responses.

Initially, subjects in C_{pct} and C_{pcnt} all received training to familiarize them with the concept of the pie-chart form of response. This took the form of a series of examples which illustrated various probabilities using pie charts. The charts were divided on their circumference into 10 equal segments. The illustrations were of different-sized segments corresponding to different probabilities - e.g. .10, .25, .60, .70, etc. (Appendices B.3 and B.4).

The subjects in C_{pct} then received the same instructions and exercise as C_{lst}, with the exception that the training in probability theory and exercise utilized the pie-chart form of response rather than the linear scale (Appendices B.3 and C). The subjects in C_{pcnt} went through the same instructions and exercise as those in C_{lsnt} to practice using this response mode for initial estimation and re-estimation of probabilities - however there was no reference to the axiomatic requirements of the Pascalian/Bayesian probability framework (Appendices B.4 and C).

The training exercise which preceded the experimental cases asked the respondents to estimate their "likelihood of guilt" rather than using the term "probability", and the experimental cases used the same terminology. Prior research in behavioural decision theory in accounting and auditing has indicated that subjects have difficulty with the term probability - a less formal term was substituted instead [Chesley 1979; Gibbins 1984].

In the demographic information provided by the subjects, almost all (over 90%) identified themselves as having graduated from a B. Comm. or B.B.A. programme where they received "training in probability theory as part of another course". Examination of the texts used in these courses at the University of British Columbia and Simon Fraser University (which provide the vast majority of Vancouver chartered accountants) showed them to be a basic introduction to two or three common probability distributions (e.g. binomial, Poisson, normal) and their applications in basic statistical inference. Based on this evidence it seems reasonable to assume that the subjects were not sophisticated with respect to probabilistic concepts and were not likely to interpret the term "likelihood" as referring to the likelihood ratio in Bayesian revision rather than simply as a synonym for "probability." [29]

Each subject received an experimental package. On the first page of the package was a brief narrative description of the client and the situation - an ongoing audit in which indications of a material fraud had been discovered. The narrative provided a brief description of what had been discovered to date and brief descriptions of five possible suspects. The subjects were informed that senior management had requested that a further investigation be undertaken to determine the extent of the fraud and the identity of the perpetrator. They were also told that the guilty party must be one of the five suspects and that there was no possibility of collusion between the suspects.

---

[29] Support for the interchangeability of the two terms is provided by a dictionary definition of probability: "in statistics, likelihood of an event" [Thorndyke-Barnhardt 1965, p. 619], and the fact that Roget's Thesaurus lists probability and likelihood as synonyms for each other [Lewis 1964, p. 290 and p. 377].

Based on the information known to date the subjects were asked to estimate their subjective likelihoods of guilt for each of the five suspects (Appendix D.1). Subsequent to this initial estimation eight additional pieces of information were presented to them one at a time, such as might be disclosed by additional investigation. (See Appendix D.2 for a list of the additional information.) The information was intended to pertain to the possible guilt or innocence of the suspects. After each piece of information the subjects were asked to re-estimate their subjective likelihoods for each of the five suspects. Each additional piece of information and request for re-estimation was presented on a separate page of the booklet, to reflect the manner in which an investigation might turn up different pieces of evidence over time. The diagnosticity of the information was subject to a manipulation check (described in section (g)).

For subjects in the C_{lst} and C_{lsnt} conditions the five suspects were listed immediately below the initial narrative (for initial estimation) and again on each page following the additional piece of evidence. The order of the list of suspects was randomized for each presentation. Below each suspect's name was a rating scale (described above) which the subjects used to estimate their subjective likelihood of guilt for that suspect. The experimental material was identical for the two groups, with the exception that at the bottom of each page requesting estimation of probabilities given to C_{lst}, was a printed reminder that the sum of the probabilities should be equal to 1.0. (See Appendix D.3 for the response format.)

For C_{pct} and C_{pcnt} the suspects were listed as for the other two groups, but beside the list was a pie-chart with the circumference marked into ten equal sections (as in the training exercise) and the centre marked with a "•". The subjects were instructed to divide the pie into segments representing their subjective likelihoods of guilt for each of the suspects and to mark the segments with the name or initials of that suspect (Appendix D.4).

All subjects were provided with a pencil and an eraser. It was emphasized that although they were not to go back and change a previous set of estimations after they had started the next set, they were free to change the "current" set as many times as they wished, until they were satisfied with it. In addition, it was verbally stressed to the subjects in C_{pct} and C_{pcnt} that if they felt the probability of a particular suspect to be zero, they were free to exclude that suspect from the pie graph.

Measurement of the subjects' responses was straightforward. For C_{lst} and C_{lsnt} it was a matter of measuring the distance from the origin ("0.0") to the slashes on the scales at each iteration. The task was made much easier because many subjects (especially in C_{lst}) used the divisions already on the scales – i.e. 10%, 20%, 30%, etc. In the case of C_{pct} and C_{pcnt} it required slightly more time. If the subject constructed his or her segments to coincide with the already-present divisions on the circumference of the graph, the proportionate size of the segments could be obtained immediately. If the subject did not make use of the divisions, a specially constructed circular template was used to obtain the size of the segments by measuring the length of the arc described by the intersection of the circumference of the circle with the rays that the subject had drawn out from the origin.

After the ninth iteration all the subjects read that it was now the end of the day and that the investigation would continue tomorrow. However, senior management was interested in knowing their impressions based on the evidence so far. The subjects were asked to indicate how confident they were about their final overall assessments. They were asked to indicate their confidence on a continuous scale anchored by "1 – very unconfident" and "7 – very confident" (Appendix D.5).[10]

(f) Manipulation Checks

In order for the analysis of subjects' probability revisions to be meaningful, it is essential that the information provided subsequent to the baseline estimation be perceived by the subjects as diagnostic. There should be evidence that one could expect these pieces of information to affect subjects' assessments of the probabilities of the alternatives. The evidence was intended to be persuasive rather than conclusive. (e.g. It would not be desirable for the second piece of evidence to prove conclusively the guilt of suspect number four regardless of any other evidence that could be forthcoming.) In order to test whether this objective was achieved, after the completion of the experiment the subjects were asked to provide information for a manipulation check by rating each of the five pieces of additional evidence on a continuous scale from "1 – irrelevant" to "7 – very relevant" in order to indicate their perception of the importance of each in relation to the assessment of guilt of the suspects. The results of the manipulation check supported the validity of the evidence. Mean values for each piece of information were computed for each condition. The mean scores within each group were compared by means of a one–way MANOVA. The results indicate that the evidence was perceived similarly by each group ($F = 1.681$, $p = .268$). The relevance scores, and changes in probability assessments were also compared. Since there was no difference in the groups' relevance ratings, the responses of all 140 subjects were combined for the comparison. The average ratings for the eight pieces of evidence were all in the middle range (from a low of 3.75 to a high of 5.65). The average change in probability assessments ranged from $|5.95\%|$ to $|16.19\%|$. Comparison of the two sets of variables showed the correlation between the relevancy ratings and the changes in probability to be significant at $p < .02$.

---

[10] This follows the construction of the confidence scale in Weber [1980].

An effect associated with the two response modes which was observed in the course of coding and transcribing the subjects' responses supports the contention that the pie chart response mode significantly reduced the amount of arithmetic processing involved for the experimental groups who employed it to make their probabilisitic estimates. Some subjects always made their estimates in multiples of 5% – i.e. every one of the 45 estimates (5 suspects X 9 iterations) ended in a zero or five – while others did not follow such a pattern.

A tabulation was constructed of the subjects who used only "even" estimates (always ending in zero or five per cent) versus the subjects who did not always use even estimates. (Strictly interpreted, this method of categorization could assign membership in the latter group to a subject who made only one "odd" estimate. However this was never the case – if a subject did use an "odd" estimate, he or she did so frequently throughout the estimation process, often as many as 15 or 20 times.) The results of this tabulation are shown in Table IV, below:

	Condition			
Type	C_{pct}	C_{pcnt}	C_{lst}	C_{lsnt}
Even	14	14	34	31
Odd	21	21	1	4

Table IV. "Even" versus "Odd" Responses

A Chi-square test produces a value of $\chi^2 = 44.42$, $p < .001$. The subjects using the linear scales were much more prone to use only "even" estimates than were the subjects using the pie charts. This is consistent with the inference that the subjects in the linear scale conditions were attempting to make their answers easier to manipulate arithmetically, while the subjects in the pie chart conditions were not as influenced by the need for arithmetic processing.

(g) Results

H_1

The results unequivocally support the contention that decision makers can be taught to make decisions within the axiomatic probability framework. Each of the 35 subjects in C_{lst} (linear scale tutored) complied with completeness and complementarity in their initial estimates and in each of the eight subsequent revisions – the sum of their probability estimates for the five suspects was always equal to 100%. By contrast, the results for C_{lsnt} (the untutored group) departed markedly from completeness and complementarity. The highest average estimate over the nine assessments was subject #17 with an average of 336% (range 310% to 360%). The lowest average estimate was subject #27 with an average of 97% (range 85% to 105%). The grand mean of the 35 subjects for the nine assessments was 205%. Table V. below shows the results for group C_{lsnt} for each of the nine iterations and Figure 2. shows a profile of the mean estimates for the nine iterations:

	Iteration (%)								
	Initial	#1	#2	#3	#4	#5	#6	#7	#8
X	201	208	216	209	215	196	211	195	196
s	81.3	78.8	72.6	73.8	75.3	73.3	76.1	69.1	69.2
low	90	90	105	95	90	90	90	80	80
high	360	365	350	330	340	330	350	340	340

Table V. Probability Estimates for C_{lsnt}

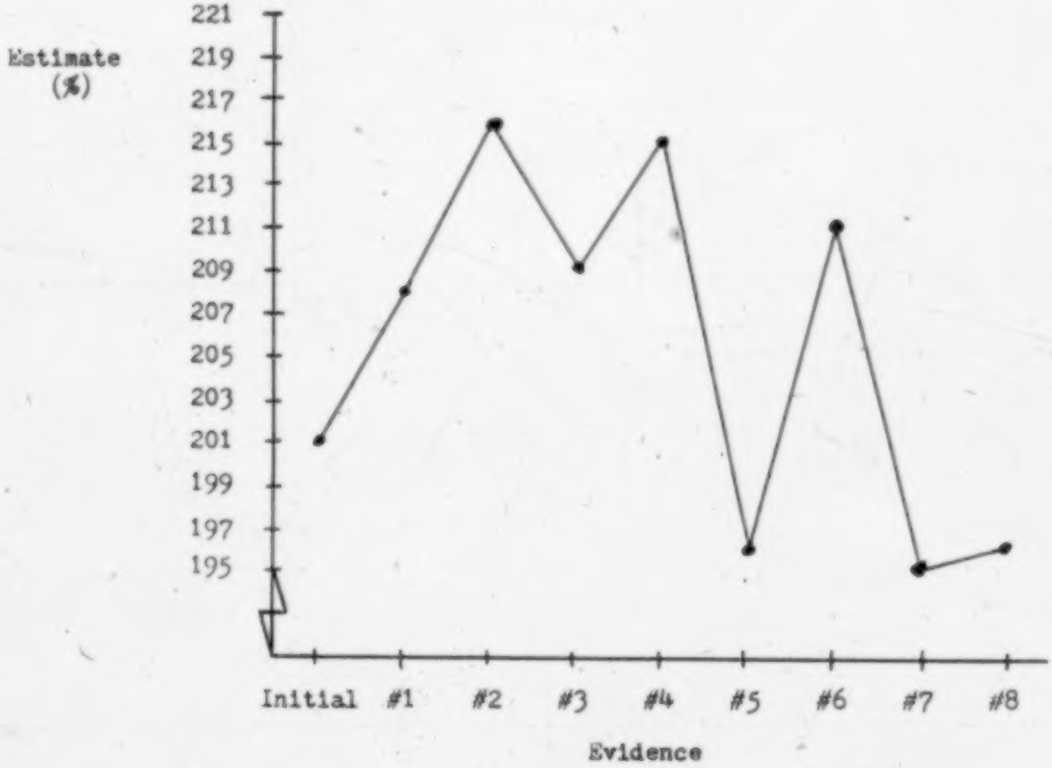


Figure 2. Profile of Mean Probability Estimates for C_{lsnt} at Each Iteration

The results of the test of H_1 can also be shown in the form of a 2 X 2 contingency table:

	Condition	
	C_{lst}	C_{lsnt}
complies with Completeness and Complementarity	35	5*
does not comply with Completeness and Complementarity	0	30

*all nine estimates of probabilities for sum of five suspects within 100% +/- 30%

Table VI. Responses to Probability Estimation Task for Groups C_{lst} and C_{lsnt}

It is clear from visual inspection of Table VI. that very many more of the subjects in the tutored group engaged in eliminations and resuscitations. A Chi-square test confirms the difference to be significant at $p < .0000$ ($\chi^2 = 52.5$).

H_2

The results provide strong support for the second hypothesis as well. Subjects who received tutoring in probability theory in order to enable them to use the linear scales to make their estimates in conformance with the axioms of probability had their information processing behaviour affected as an unintended side-effect of that training. Specifically, of the 35 subjects in group C_{lst}, 28 (80%) engaged in resuscitations, while only 7 (20%) did not. The number of times those 28 subjects engaged in resuscitations ranged from 1 to 7 with an average of 1.5. In marked contrast, none of the other subjects in any of the other three groups engaged in any resuscitations. Table VII. below summarizes the results by group:

	Condition			
Pattern	C_{pct}	C_{pcnt}	C_{lst}	C_{lsnt}
resuscitations	0	0	28	0
no resuscitations	35	35	7	35

Table VII. Frequency of Resuscitations by Group

A Chi-square test of the frequencies in Table VII. produced a value of $\chi^2 = 105$, $p < .0000$. There is a vast difference between the behaviour of C_{lst} and the behaviour of the other three groups. As hypothesized, the subjects using the pie charts did not engage in resuscitations. (It is interesting to note that the subjects in C_{lsnt}, who also used the linear scales, did not engage in any resuscitations either. This is a more extreme result than RH, who did observe a few in their experiment.)

Analysis of the responses in C_{lst} shows that while the resuscitations involved some of the suspects more often than others, all of the suspects experienced some incidence of the same phenonenon. The most frequently eliminated and resuscitated suspect was John Sims, chief warehouseman. The first piece of evidence about this suspect was negative (it was intended to reduce the subjects' belief in his probable guilt). However, the first piece of evidence concerning Susan Evans, salesperson, was positive and she was the second most frequently eliminated and resuscitated suspect. (See Appendices D.1 and D.2 for descriptions of the suspects and the additional evidence respectively.)

Summary of H_1 and H_2

The results of the tests of these two hypotheses strongly support the findings of RH. It is possible to teach subjects to conform to the probability axioms in their probability estimations, as a comparison of the results of the subjects in C_{lst} and C_{lsnt} clearly shows. However there is a "cost" to this conformance (in addition to the resources used in training the subjects). The tutored subjects resorted to frequent premature eliminations and resuscitations of different suspects throughout the completion of the case.[31]

The contrasting performance of subjects in groups C_{pct} and C_{pcnt} is illuminating. With regard to

---

[31] There were instances of eliminations of various suspects in all groups. However, eliminations of suspects, even early eliminations, are normatively acceptable, as the diagnosticity of any piece or pieces of evidence was up to the individual. It is only eliminations and subsequent resuscitations which are in contravention of axiomatic probability and are by definition "premature".

H_1, tutoring in probability theory was not necessary to ensure the adherence of the subjects in C_{pcnt} to completeness and complementarity – the simple pie chart decision aid did that. With regard to H_2, while prior tutoring in probability theory had adverse consequences for the subjects in C_{lst}, it had no such effect on subjects in C_{pct}, who showed no evidence of resuscitations. Thus it could be argued that in order for judges to follow the basic principles of axiomatic probability theory in their uncertain decisions, tutoring is neither necessary nor sufficient. A superior approach is to provide judges with a simple decision aid which removes the need for arithmetical processing. Their intuitive information processing behaviour (i.e. without the intervening variable of arithmetic calculation) is consistent with the framework of axiomatic probability.

H_3

H_3 predicted that the subjects in the four different treatment conditions would display different levels of agreement within themselves regarding their final estimates of guilt of the five suspects. Analysis of the results provided qualified support for H_3.

As a first step, histograms of the five final probability estimates (one for each suspect) were constructed for each of the four experimental conditions. None of the histograms showed drastic departures from a normal distribution, and the kurtosis and skewness statistics were all within reasonable parameters. There were no outliers. A series of 20 Kolomogorov–Smirnoff tests for normality (5 suspect estimates per treatment condition X 4 treatment conditions) confirmed that there were no systematic departures from normality in any of the groups' responses. However, non–parametric procedures were deemed more appropriate due to the nature of the data matrices – although visual inspection and descriptive statistics indicated that the probability estimates were reasonable approximations of normal distributions, the fact that the sum of the estimates was constrained to equal 100% each time was clearly in violation of the assumptions necessary for reliance on parametric procedures.

Simple averages of the standard deviations of the five mean estimates (one for each suspect) within each group showed that C_{lst} had the highest average variation ($s = 11.5$), while the other three were clustered together at a lower value ($s = 9.5$). The significance or non-significance of any differences cannot be validly inferred from a comparison of averages, but it does seem that there is some difference between them.

Pairwise correlation matrices were computed using Spearman's Rank Order Correlation procedure for the five final probability estimates of the 35 subjects in each group [$n(n-1)/2 = (5 \text{ X } 4)/2 = 10$ elements per matrix]. The correlations showed some differences between groups. The average values of $|\rho|$ were as follows: $|\rho|_{pcnt} = .4237 > |\rho|_{pct} = .3400 > |\rho|_{lst} = .2983 > |\rho|_{lsnt} = .2908$. (The distributional properties of the average correlation coefficient are not well known, but for purposes of comparision coefficients of .3361 and .4418 are required for attained significance levels of .05 and .01 respectively for pairwise comparisons.) This result seems to be consistent with the general expectation that the pie-charts would result in greater within-group agreement, but again the significance of the differences is not determinable.

A final non-parametric procedure was to calculate Kendall's Coefficient of Concordance [Hays 1981, pp. 606–8], using each condition's final estimates. The statistic, W, is a measure of overall rather than pairwise agreement in rankings within treatment conditions. The values of W_i, $i = 1,\ldots,4$ were all significant at $p \leq .02$ which indicates that although the agreement within the pie chart conditions may have been slightly higher by some measures, the degree of within-group agreement on the final probability of guilt of the five suspects was high in all conditions.

Parametric procedures were performed on the first four of the five final probability estimates (to alleviate the problem of linear dependence) and also on all five. One-way MANOVA's were done to compare the final estimates across the four conditions (e.g. a 4 X 1 MANOVA on a 5-variate/4-variate response vector.) The results supported the equality of the four mean response vectors (five out of five final estimates – $F = 1.400$, $p = .176$; four out of five final estimates – $F = 1.169$, $p = .084$). While Box's test for the equality of co-variance matrices [Cooley and Lohnes 1971, pp. 228–29] did not support equality for either the 5 X 5 matrix ($F = 4.819$. $p < .05$) or the 4 X 4 matrix ($F = 3.341$, $p < .05$), Seber points out that the test is "...overly sensitive to departures from H_0 in that it detects differences among the dispersion matrices that have little effect on the test for means [Seber 1984, p. 449]." Thus some reliance may be placed on the results of the MANOVA. Figure 3. on the next page presents profiles of the final mean estimates for the five suspects for each of the four experimental conditions. It can be seen from an examination of Figure 3. that the final mean estimates of the probability of guilt of each of the five suspects was very similar across the four experimental conditions.[32] [33]

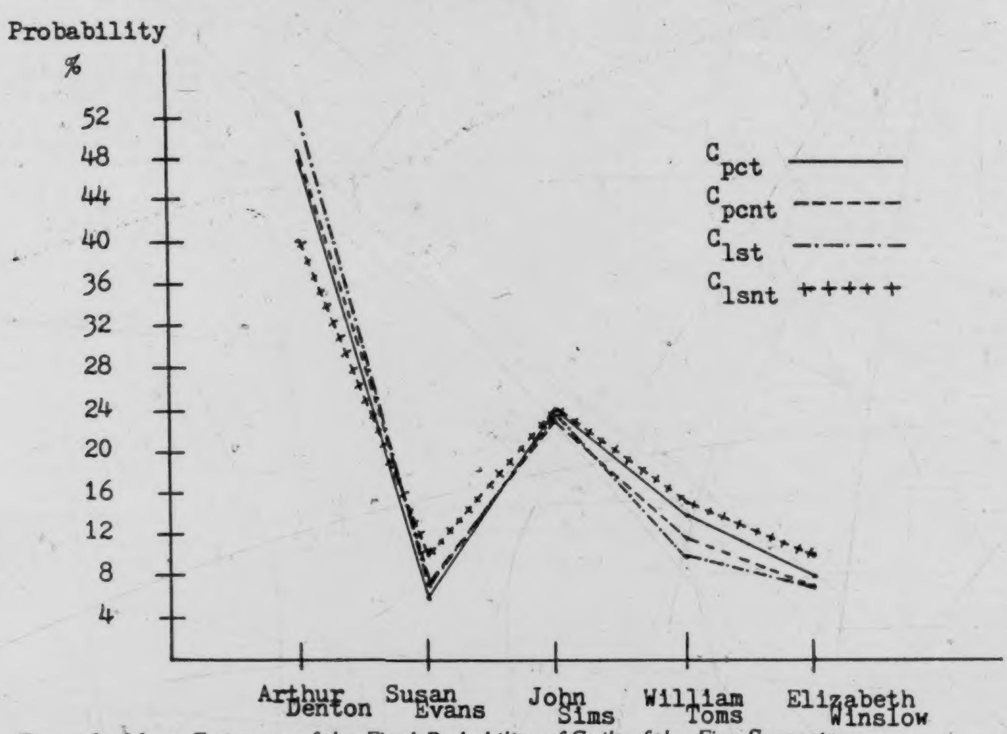


Figure 3. Mean Estimates of the Final Probability of Guilt of the Five Suspects

---

[32] There are several factors which may have contributed to the similarity of the estimates across the four groups. First, the values within each treatment condition are averages, within those averages there are wide individual variations in the estimates for each of the suspects. Second, the fact that the five estimates were constrained to sum to a constant (e.g. 1.0 or 100%) may have "forced" some degree of convergence. Finally, it is also possible that the clues were perceived as more diagnostic than expected, although the clue ratings in the manipulation check did not indicate an extremely high degree of diagnosticity. This result may have weakened the test of H_3 slightly, but since the hypothesis is concerned with the variance of the estimates it seems reasonable to suggest that the effect is not strong.

[33] The estimates of subjects in C_{lsnt} have been scaled *ex post* to satisfy $\Sigma P(A_i) = 100\%$.

H_4

H_4 concerns the confidence of the subjects in their final overall probability assessments. The order of the mean confidence scores was C_{pcnt} (X = 4.654) > C_{pct} (X = 4.643) > C_{lsnt} (X = 4.214) > C_{lst} (X = 3.460). (See Appendix G.1 for summarized statistical results.)

A 2 X 2 ANOVA (response mode X tutoring) on the confidence scores showed that the response mode (pie charts vs. linear scales) was highly significant (F = 16.84, < .0001). The tutoring variable was almost significant (p = .055), as was the interaction of the two (p = .062). This supports the inference that the visually-based pie chart response mode does have an effect on some of the secondary characteristics of information processing behaviour. It is consistent with the idea that an advantage of the pie chart form of response mode is that it stresses the interrelationship of the subcomponents to the whole and allows the judge to consider all aspects of the judgment. It provides, in the vernacular, a more "holistic" approach. It also suggests that tutoring in probability theory does not lead judges to have any more confidence in their subsequent probabilistic judgments and may in fact be a distractor.[14]

Post hoc comparison of the means was done by Newman-Keuls procedure. (Preliminary tests showed that the variances of the confidence scores of the four groups could be considered equal (Box's test: F = .160, p = .9219; Levene's test: F = .270, p = .8496 [Brown and Forsythe 1974b]). The test showed that C_{pct}, C_{pcnt} and C_{lsnt} were not significantly different and that C_{lst} was different, at a significance level of p < .01, from each of the other three. This was supported by pairwise t-tests on the means of the four groups, which showed the difference between C_{lst} and each of the other three to be significant at .05 even if allowance was made for the reduced protection level afforded with multiple comparisons.

---

[14] The relationship of confidence to judgment quality is not clear. The two have been linked in the auditing literature (e.g. Weber [1980]) and Shanteau [1989] suggests that confidence is a necessary condition for expertise, but the testimony literature has not found a positive correlation between confidence and accuracy. Since the design of this experiment precluded direct measures of accuracy, these results do not speak to this issue directly. However, it does appear that consensus and confidence are positively correlated. With respect to the SEU model, it is suggested that an increased level of confidence in a decision may be an intangible factor which contributes to increased utility to the decision maker.

H_5

The fifth hypothesis predicted differences in the perception of the difficulty of the probability estimation task using the different response modes X different levels of instruction. (See Appendix G.2 for summarized statistical results.) A 2 X 2 ANOVA on the difficulty scores showed significant differences by response mode ($F = 23.46$, $p < .0001$) and by level of tutoring ($F = 4.44$, $p = .0369$). The interaction the two was also highly significant ($F = 6.23$, $p = .0137$). The order of the mean difficulty scores was C_{pct} ($X = 3.051$) < C_{pcnt} ($X = 3.140$) < C_{lsnt} ($X = 3.674$) < C_{lst} ($X = 4.723$). These results support the inference that the pie chart response form is perceived as easier to use than the linear scales, and also that tutoring the subjects and requiring them to use the linear scales in conformance with the axioms resulted in this condition being perceived as the most difficult.

Using Neumann–Keuls, it was again found that the mean difficulty score for C_{lst} was statistically different from the other three at $p < .01$. The other three were not significantly different. This again suggests that it was not so much the linear scales that the subjects perceived to be difficult as it was the combination of linear scales and axiomatic probability. (Although Box's [1953] test and Levene's test both rejected the homogeneity of variances among the four groups at $p < .05$, a one–way analysis of variance with the variances not assumed to be equal (Brown–Forsyth test [Brown and Forsyth 1974a]) showed a between–groups significance level of $p < .0001$., and Boneau [1960]; Donaldson [1968], and Tiku [1971] have presented empirical evidence which supports the robustness of the F–test.) Pairwise t–tests indicated a significant difference between C_{lst}, and C_{pct} and C_{pcnt}, even allowing for the reduced protection level afforded by the multiple comparisons. These results are consistent with the predictions of H_5.

Longitudinal Analysis

In order to examine the process by which the subjects arrived at their final judgments, a number of non–parametric and parametric tests were performed over the full nine iterations. (See

Appendix G.3 for summarized statistical results.)

The first parametric procedure was to generate five (one for each suspect) Profile Analyses [Fox and Guire 1976, p. 130; Johnson and Wichern 1982, pp. 263-68]. The Profile Analyses compared the 9-variate vectors of mean probability estimates (initial plus eight subsequent items of evidence) across the four conditions. The profiles provide a "map" of how the subjects arrived at their final probability estimates. Visual examination of the profiles of the four conditions over the 9 iterations for each of the five suspects showed that they were not widely divergent across conditions for any of the suspects, but neither did they appear to be that "close".

Statistically, Profile Analysis decomposes the overall question "Are the population mean vectors the same?" into three sequential questions – 1. Are the profiles parallel? 2. If the profiles are parallel, are they coincident? 3. If the profiles are coincident, are they level? (The first "no" answer terminates the sequential process.) In order to answer these questions, Profile Analysis employs contrast matrices with the pooled covariance matrix to test if the experimental interventions have the same (average) effects on the different conditions. In the case of the present study, the focus was on the first two questions above. The results of the manipulation check regarding the diagnosticity of the items of evidence would strongly suggest that the profiles would not be level. (Indeed, if they were level, that would be evidence that the experimental manipulation did not work.)

The results were that in each of the five analyses, the null hypothesis of parallelism of the profiles was rejected at $p < .05$. Since the hypothesis of parallelism was clearly rejected , the hypothesis of coincidence is automatically rejected as well.[35]

---

[35] It must be noted that all the assumptions underlying this method of analysis were not fully satisfied. Although the data met the conditions of normality, as evidenced by a series of Kolomogorov-Smirnoff tests for normality, diagnostic tests also showed that the assumption of equality of covariance matrices was not well met. However, the primary concern in this case was the false acceptance of the null hypothesis of parallelism (or coincidence if applicable) of estimate vectors rather than the false rejection of the null hypothesis. Visual inspection of the profiles strongly indicated that the results of the statistical tests were valid.

Visual examination of the profiles in each of the suspect/cases indicated that the most "different" of the profiles was consistently C_{lst}. This inference was supported by the results of two additional procedures. First, the Profile Analyses were redone, omitting the C_{lst} condition. In the cases of suspects Arthur Denton, John Sims and Elizabeth Winslow, the hypothesis of parallelism was not rejected, and the other two suspects while still significantly not parallel, did show a noticeable reduction in significance.[16]

As another indication of the similarity or difference among the profiles, a comparison was made of the frequency of interactions (defined as crossing of the profiles) between the conditions. In total (over the five analyses) C_{pct}, C_{pcnt} and C_{lsnt} interacted among themselves 18 times, while C_{lst} interacted with the other three 31 times, the difference being significant by a p-test at $p < .025$. The more erratic behaviour of C_{lst} is consistent with the explanation that the estimates of the subjects in this condition were being determined at least in part by the arithmetical constraints imposed on them.

Another statistical test which was used to gain insight into the process by which subjects reached their final judgments was to analyze the experiment as a completely crossed 2 X 2 Repeated Measures design. The advantage of the Repeated Measures approach over Profile Analysis is that it allowed examination in more detail of the differences between the response vectors of the four conditions. The evidence from Profile Analysis and the related tests (see preceding paragraphs) indicated that C_{lst} was the most "different" condition. The purpose of this analysis was to study somewhat more closely, the factors which may have contributed to that difference.

A set of five (one for each suspect) 2 X 2 (response mode X prior tutoring) Repeated Measures analyses were done [Bock 1975, Ch. 7; Dixon 1983, pp. 362-69] on the 9 X 1 vectors of probability estimates across the four conditions.

---

[16] The omission of C_{lst} did not result in any of the sets of three remaining profiles being statistically coincident. Any causal explanation for this result is not apparent and is a question for future investigation. No condition was dominated by position, experience or firm effects.

A major advantage of the Repeated Measures design is its potential for greater efficiency. Ordinarily, the error term for testing the significance of main and/or interaction effects contains an element of variability due to the inherent variability of each subject's response. Because in this experiment each subject made nine successive probability estimates, this variation in responses due to random influences could be extracted from the error term. This should allow the error term to be smaller and should lead to more efficient tests of the effects of interest.[37]

In each of the five analyses the results showed a significant difference in the pattern of estimates across the four conditions. As well as clearly indicating the inequality of the 9 X 1 estimate vectors in each case the analyses examined the patterns for evidence of interactions of the estimates with shape (pie graph versus linear scales - "ET"), with prior tutoring (yes versus no - "ES"), and a three way estimate-shape-tutoring interaction ("EST"). The only evidence of a consistent interaction was ET, which was significant at $p < .05$ three out of five times.

This could be interpreted as indicating that prior tutoring in probability theory versus no prior tutoring was a factor in the process by which the subjects arrived at their final probability estimates. This is consistent with expectations for the responses of subjects using the linear scale response form, but not an expected result for subjects using the pie chart response form. However, the results of the Profile Analyses and the non-parametric tests discussed in the preceding paragraphs suggest that the significance of the ET interaction is mostly due to the influence of C_{lst}.

---

[37] The additional restriction which this imposes on the data relates to the correlation between the paired responses of any subject (e.g. y_i, y_j). It is the correlation between the responses which allows the reduction of the error term, but the validity of the Repeated Measures procedure also requires the additional assumptions of equal covariances and equal correlation coefficients. Tests of these assumptions, known in combination as the "sphericity assumption" resulted in rejection at a high level of significance in each of the five analyses. Although there is support in the literature for the robustness of the F-tests with respect to both Type I and Type II errors [Gabrielsson and Seeger [1971]; Stoloff [1970] and the analysis reports the results of two conservative F-tests (due to Huynh and Feldt [1979, 1980] and Greenhouse and Geisser [1959]) which adjust for this condition by reducing the degrees of freedom contributed by the within factor, the results of the Repeated Measures analysis is best interpreted as further persuasive evidence consistent with the results of the prior tests.

Further support for the inference that the tutoring versus no tutoring factor had more of an effect on the subjects using the linear scales can be seen by first comparing the difference in the frequency of interactions between C_{pct} and C_{pcnt} versus between C_{lst} and C_{lsnt} and then comparing the difference in the frequency of interactions between C_{pct} and C_{lst} versus between C_{pcnt} and C_{lsnt}. The difference in the frequency of interactions (defined as crossing of the profiles) between C_{pct} and C_{pcnt} (3) and the frequency of interactions between C_{lst} and C_{lsnt} (14) was significant at $P < .01$. Combined the other way, the difference in the frequency of interactions between C_{pct} and C_{lst} (10) compared to the number of interactions between C_{pcnt} and C_{lsnt} (6) was not significant at $p < .05$.

Thus, it appears that although the four conditions arrived at essentially the same final judgments concerning the probability of guilt of the five suspects, they took slightly different "routes" to get there. However, although there were differences in responses among the four conditions over the nine iterations, those differences were somewhat erratic. There was little evidence of really strong systematic trends or patterns. The statistical significance of any differences for the most part has no clear practical interpretation. There is some indication, consistent with previous evidence, that the combination of quantitative linear scales and prior tutoring in axiomatic probability theory affected the responses of the subjects in C_{lst} in the direction of increased variability over the nine iterations.

Summary

Overall, the results of the Fraud Investigation experiment indicate that with minimal prompting, people's intuitive information processing behaviour does conform to Pascalian/Bayesian notions of probability. The results also suggest the superiority of the pie chart form of response mode as a means of eliciting probabilistic estimates on the basis of qualitative evidence. The pie chart response mode eliminates the probability axiom violations, does not result in premature eliminations and resusciations, and appears to have beneficial consequences in terms of facilitating the judgments themselves. This supports the notion that to improve the ability of judges to assess uncertain evidence probabilistically, instruction in normative probability theory is not the most efficacious method. It is neither "necessary nor sufficient". The simple ex ante scaling device of having judges make their estimates graphically using a circle graph cures the violations and "irrational" decision behaviour, is perceived as easier to use, and results in increased within-group consensus and increased confidence in the final judgments.

VI. THE INTERNAL CONTROL RISK EVALUATION EXPERIMENT

(a) Background

The second experiment was set in the context of internal control evaluation - one of the most important judgments that an auditor must make. The internal control system is the system that the client has put in place to prevent or detect accidental or deliberate errors in the processing of the firm's transactions. It is the auditor's responsibility under Generally Accepted Auditing Standards (GAAS) to evaluate the client's system of internal control [CICA Handbook 1986/1965, sec. 5100.02]. Based on his or her judgment about the risk of the internal control system, the auditor must select a level of substantive testing (i.e. the intensity of his or her own scrutiny of the client's financial records) in order to obtain reasonable assurance that the financial statements do not contain material errors or misstatements. However, the manner in which the results of this internal control judgment are to be incorporated into decisions about substantive testing is not discussed in the Handbook. One of the purposes of this experiment is to examine the effect on subsequent substantive testing decisions, of different forms of internal control risk evaluation - an explicit (i.e. written) judgment in two probabilistic modes and one non-probabilistic mode, and an implicit (i.e. no requirement for a written judgment) mode.

One reason for examining internal control evaluation within the framework of the Probabilistic Judgment Paradigm is that the nature of the ultimate response required may tend to obscure the intervening decision process. The auditor ultimately must choose a particular course of action - a level of substantive testing. The connection between the initial decision situation and the final decision may appear to be deterministic [Gibbins 1984] even when the decision itself has a probabilistic component, because that component is not observable. The subsequent rationalization and justification of the choice, which may be intended in part to minimize or mask any uncertainty, could further obscure any probabilistic component to this process. (See Emby and Gibbins [1988] for evidence on the importance of justification in audit decision making.)

The auditor obtains and evaluates evidence (often uncertain and incomplete) about the strength of the client's system of internal control. It may be that what the auditor does is use this evidence to evaluate the probability that the internal control risk is "low, medium or high", and based on this probabilistic estimate, go on to select a level of substantive testing. The experiment contrasts two types of probabilistic internal control evaluation with a non-probabilistic mode in order to examine internal control risk assessment.

(b) Experimental Design

Experiment two contained some of the same elements as experiment one, but the difference in focus between the two experiments necessitated differences in design. The experimental design was a 4 X 1 factorial. Two of the experimental conditions (C_{lst} and C_{pct}) were carried over from experiment one. The tutored linear scale condition was replicated from experiment one in order to observe the presence or absence of resuscitations.[38] The tutored pie chart condition was replicated in experiment two in order to facilitate a comparison of the linear scale response mode and pie chart response mode without the potential confound of differing levels of prior training.[39] The subjects in the two groups were asked to make a series of probabilistic judgments about the level of internal control risk in a payroll system.[40] There were four levels - "extremely low", "low", "medium", and "high".[41]

---

[38] It was expected, prior to the experiments, that if resuscitations occurred, it would be primarily in the linear scale tutored group rather than the untutored group. The results of experiment one confirmed that expectation.

[39] It was expected, prior to the experiment, that in the pie chart conditions, "tutoring versus no tutoring" would not have a significant effect. The results of experiment one confirmed that expectation.

[40] Most previous research into auditor judgment with respect to internal control has examined auditors' decisions concerning internal control *strength*, which is the complement of internal control risk. However, because of the link between internal control risk, substantive testing risk, and appropriate level of substantive testing, this research focuses on the intermediate assessment of internal control risk.

[41] An important premise is that the subjects perceived the alternatives to be mutually exclusive and exhaustive. For that reason, a brief explanation of what would and would not be included in each of the four alternatives was provided verbally in the

In addition, another group of subjects, designated C_{nps}, used a non-probabilistic continuous scale anchored by "1 – extremely low" and "7 – high" to record their internal control judgments.[42] A fourth group, designated C, were provided with the same items of evidence but were not required to make explicit assessments of internal control risk before choosing a level of substantive testing.

The subjects in C_{lst} and C_{pct} were carried over intact from experiment one. The subjects in groups C_{lsnt} and C were reassigned randomly from C_{pcnt} and C_{lsnt} in experiment one.

The auditors' judgments concerning the appropriate amount of substantive testing were elicited on a continuous scale from "1 – minimal" to "7 – exhaustive".[43] The subjects were asked to make this judgment on the basis of the initial narrative and again after receipt of five additional pieces of evidence about the internal control system. A comparison of the judgments of these four groups concerning the appropriate level of substantive testing allowed inferences concerning the relative efficiency and/or effectiveness of the different modes of decision making. They were also asked to express their confidence in their initial and final judgments on a continuous 7-point scale ranging from "1 – not confident" to "7 – very confident", and rate the difficulty of the risk estimation/substantive testing level task, similarly to experiment one. The experiment was of the type described by Campbell and Stanley [1966, pp. 13–24] as a pre-test, post-test design. Figure 4

---

[41](cont'd) pre-experimental training.

[42] This group's evaluations of control risk were representative of procedures which are used currently in public practice by a number of CA firms. The scale is a "synthesis" in that the documentation of each of the firms is different – one uses a three-valued verbal scale, one uses a discrete quantitative scale, one uses a continuous quantitative scale, etc. Some are based on evaluations of internal control strength and some are based on evaluations of internal control risk (essentially "strength" = "1 – risk" or vice versa).

[43] The word "minimal" rather than "none" was chosen as an endpoint of the scale for two related reasons. First, Generally Accepted Auditing Standards have historically been interpreted as precluding the auditor from eliminating substantive testing altogether. As a result the effect of using "none" as the minimum endpoint would be to compress the effective scale to something less than a seven-point scale and introduce bias into the responses (see Emory [1980, pp. 263–64] for discussion of this potential problem).

presents a schematic diagramme of experiment two:

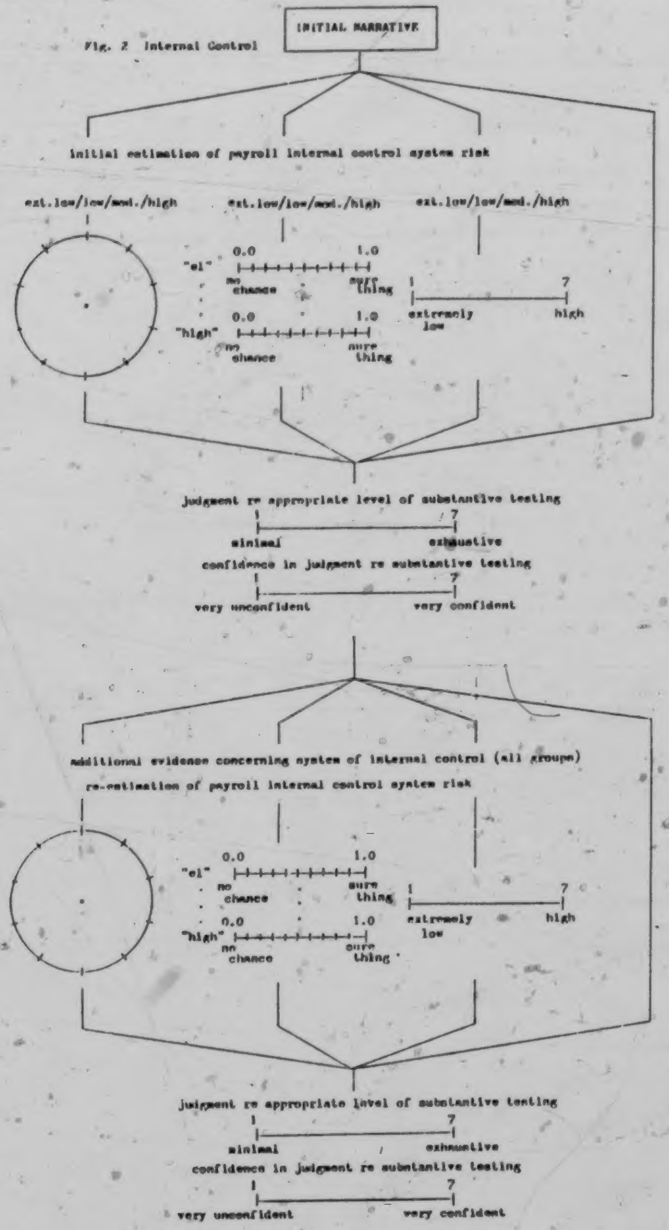


Fig. 2 Internal Control
INITIAL NARRATIVE
initial estimation of payroll internal control system risk
ext.low/low/med./high
ext.low/low/med./high
ext.low/low/med./high
0.0
1.0
"el"
no chance
sure thing
0.0
1.0
"high"
no chance
sure thing
1
7
extremely low
high
judgment re appropriate level of substantive testing
1
7
minimal
exhaustive
confidence in judgment re substantive testing
1
7
very unconfident
very confident
additional evidence concerning system of internal control (all groups)
re-estimation of payroll internal control system risk
0.0
1.0
"el"
no chance
sure thing
0.0
1.0
"high"
no chance
sure thing
1
7
extremely low
high
judgment re appropriate level of substantive testing
1
7
minimal
exhaustive
confidence in judgment re substantive testing
1
7
very unconfident
very confident

Figure 4. The Internal Control Risk Evaluation Experiment

(c) Expectations and Hypotheses

The expectations in this experiment relate to the effect of different modes of evaluation of internal control risk on the actual assessments of risk and on subsequent decisions regarding the appropriate extent of substantive testing.[44] The first hypothesis to be tested concerned the effect of different modes of probabilistic processing on the judgments of internal control risk. It was expected that here, as in experiment one, the need for the subjects in C_{lst} to engage in arithmetic processing would interfere with their information processing behavior. This could lead to responses on the part of these subjects which were driven more by the arithmetic than by evaluation of the evidence. This hypothesis may be stated as:

> H_6: The level of agreement among the subjects in condition C_{pct} regarding the assessment of the probabilities of the different levels of internal control risk will be higher than C_{lst}.[45]

The next two hypotheses relate to the possible effects of different modes of processing on subsequent decisions regarding the appropriate level of substantive testing. The first of these represents an extension of H_6. With respect to C_{lst} and C_{pct}, if some subjects show more agreement in the initial phase of the process (i.e. the internal control risk evaluation phase), one might expect greater agreement on the subsequent phase of the process. The inclusion of C_{nps} represents the current practice in the profession (at least among the larger firms). C compares the effects of prior explicit internal control evaluation versus no prior explicit control evaluation on the subsequent decision. It is the expectation in this experiment that one of the benefits of formal internal control evaluation and documentation is that it leads to greater agreement on the appropriate next step in the process.

---

[44] According to the matrix proposed by Felix and Niles [1988], this study would be characterized primarily as "re-evaluation X practice description." However, it is concerned with auditors' direct use of probabilistic concepts as well, and as such could also be characterized as "re-evaluation X theory testing" (p. 46).

[45] Recall that this is a two way comparison only as condition C_{nps} does not make probabilistic assessments of control risk and condition C does not make any explicit assessments at all.

Following the same reasoning as in experiment one it is possible to suggest a partial ordering of the responses. Expectations regarding the benefits of explicit risk evaluation imply that the subjects using the non-explicit response mode (C) will display the least within-group agreement. Expectations regarding the response of the other three groups are less unambiguous. The order may be $C_{pct} > C_{lst} > C_{nps}$, which could be interpreted as indicating that an intermediate decision aid does help, that a probabilistic framework is consistent with the subjects' information processing behaviour, and that the pie chart response mode does enhance the judgment process. Or it may be that $C_{nps} > C_{pct} > C_{lst}$ which could be interpreted as indicating that while an intermediate decision aid does help, the framework of probabilistic processing lessens the beneficial effects of the decision aid, especially within the constraint of the linear probability scales. Clearly, there are other possible orders, which support other inferences, but it would be premature to make wholesale predictions without stronger theoretical or empirical support. If the initial hypothesis test does reveal significant differences, *post hoc* tests can be used to identify any order effects, as was done in experiment one. This hypothesis may be stated as:

> H_7: The level of agreement among the subjects in the non-explicit condition regarding the appropriate level of substantive testing will be lowest.

Similar to experiment one, the subjects in each condition were asked to indicate their confidence in their final judgments and the mean score within each condition was compared. It was expected that the mean confidence levels across the four groups would differ – primarily with respect to a lower mean level of confidence in group C. The relative confidence levels of the other three groups may differ also, but even if the results of H_7 and H_8 indicate systematic differences between the response modes, it is more tenuous to extend this "one more step". Consequently, there are no strong expectations about any directional differences between the three conditions which require explicit judgments of internal control risk. This expectation may be stated more formally as:

> H_8: Confidence in the decision regarding the appropriate level of substantive testing will be lowest in condition C.

The final hypothesis in this experiment again relates to the perceived difficulty of using the various modes of intermediate processing. Drawing on the theoretical arguments of Gnanadesikan [1980] and Torgerson [1958], it is expected that the probabilistic estimation of internal control risk using a series of "0.0 - 1.0" linear scales while being constrained to adhere to completeness and complementarity, will be perceived as more difficult. The perceptions of the subjects in the other two experimental conditions, C_{xct} and C_{nps}, are expected to be that the level of difficulty is lower.

The question of perceived degree of difficulty can have practical ramifications. At present, firms use a deterministic approach to the evaluation of internal control and the subsequent choice of level of substantive testing. However, the probabilistic approach has merit in that it stresses the uncertainty inherent such a decision, and the importance of the judgmental nature of the decision. If the Probabilistic Judgment Paradigm model of this process yields qualitatively appropriate results, and is perceived by the subjects as relatively easy to use, it can be supported on pragmatic as well as conceptual grounds.[46] Expectations regarding the perceived difficulty relate to the increased cognitive effort required of the subjects in C_{lst}. These expectations may be stated as follows:

H_9: The difficulty ratings of the subjects in group C_{lst} will be greatest.[47]

(d) Subjects

The subjects were the same 140 practicing auditors from 20 CA firms in Vancouver.

---

[46] The preferred approach in actual professional practice may also be a function of the kind of documentation CA firms wish to have in their files as evidence of their review and evaluation of internal control leading to the choice of a level of substantive testing. However, this may involve questions of legal liability, and this study does not try to address that issue.

[47] Recall that condition C does not employ this question.

(e) Materials and Administration

The Internal Control Risk Evaluation experiment followed directly after the Fraud Investigation experiment in the experimental packages. The subjects proceeded to the second experiment at their own pace, after they had finished the Fraud Investigation experiment.

The first pages contained a brief narrative, describing a hypothetical manufacturing company (Appendix E.1). It introduced the company in general terms and provided a description of some basic structural aspects of the company's payroll internal control system.[41] Based on this initial description of the payroll system, the subjects in all conditions except C were asked to circle the category of internal control risk which in their opinion best matched the internal control system as described. (Recall that the pre-experimental training included an explanation of each of the four categories ("extremely low", "low", "medium" and "high") to ensure that they were interpreted consistently by the subjects.) The purpose of this initial categorization was twofold - first to introduce the categories which the subjects would use, and second, to reinforce the range of categories represented in the quantitative assessments. The subjects in C_{pct} used a pie chart to indicate the probability of each of the four categories, the subjects in C_{lst} used linear scales to indicate the probability of each of the four categories, and the subjects in C_{nps} used a single continuous non-probabilistic "risk" scale anchored by "1 - extremely low" and "7 - high" (Appendix E.1).

Based on the initial description of the payroll system, all subjects, including group C, were asked to make a judgment as to the appropriate amount of substantive testing and to indicate their confidence in that judgment. The subjects were then presented with additional facts, one at a time, concerning the payroll internal control system. After each fact, the subjects in conditions C_{pct}, C_{lst} and C_{nps} made new assessments of the internal control risk, using the same response modes as they used in the initial assessments. There were five additional facts - each one presented on a separate page of the booklet. The additional items of information were chosen from those identified in R. Ashton [1974a]; Ashton and Brown [1980] and Haskins [1987]. The subjects in condition C received the same pieces of information, each on a separate page as in the other conditions, but were not required to make intermediate assessments of control risk. After the fifth piece of evidence (and sixth assessment for the experimental conditions) all subjects were again asked to make a judgment about the appropriate level of substantive testing, and indicate their confidence in that judgment.

---

[41] The payroll system was chosen for this experiment for two main reasons. First, because the review of internal control is a very important step in the audit of this major system, and second, the characteristics of an adequate payroll system should be familiar to these subjects (see Ashton [1974a]; Ashton and Brown [1980]).

The final pages of the experimental package asked the subjects for the demographic data described in section (e) of the Fraud Investigation experiment.

(f) Manipulation Checks

Manipulation checks were performed to ensure that the pieces of information were perceived as diagnostic. Similar to experiment one, the subjects were asked to rate each piece of evidence in order to indicate their perception of each in relation to the question of internal control risk. The ratings were again all in the middle range, indicating that the pieces of evidence were relevant but that no single piece dominated. (The highest rated pieces of evidence were those that concerned separation of duties – confirming previous findings.) The ratings were compared across groups and the ratings were compared with changes in probability estimates.

As in the first experiment, the results of a 4 X 1 ANOVA on the relevancy ratings failed to reject the hypothesis of equality of the four mean vectors. The average correlation co–efficient between the mean relevancy ratings and the magnitude of the changes in probability estimates (i.e. $|r|$) was significant at $p < .05$.

The time required by the subjects to complete the two cases was recorded for conditions C_{pct} and C_{lst}. (This was done unobtrusively – knowledge that they were being "timed" might have caused some subjects to experience performance anxiety and resulted in biased responses. This was not difficult as the experiment was administered to relatively small groups of subjects at any one time.) If using the linear scales and being required to conform to completeness and complementarity is more cognitively complex than using the pie–charts, it could be expected that the subjects in the tutored linear scale condition would take longer, on average, to complete the experiment, than would the subjects in the tutored pie–chart condition. (Since the other subjects did not use the same response formats in the two successive experiments, the interpretation of the time it took them to complete both experiments would be somewhat problematic.) There was a highly significant difference between the average length of time required to complete the two experiments for the subjects in C_{pct} ($X = 46.6$ minutes) versus the subjects in C_{lst} ($X = 54.6$ minutes). A one–sided t–test produced a value of $t^{*} = 5.68$, $p < .001$. This is consistent with the results of H_5 and H_9, both of which showed the difficulty ratings for C_{lst} to be significantly higher that for C_{pct}.

Similar to experiment one, a tabulation of the frequency of "odd" versus "even" responses indicated that the subjects in the linear scale conditions were much more prone to use probability values that were easier to manipulate arithmetically ($\chi^2 = 36.6$, $p < .001$).

(g) Results

The responses of the subjects in C_{pct} and C_{lst} displayed similar patterns to the Fraud Investigation experiment. All subjects in both groups complied with completeness and complementarity. There were no resuscitations in C_{pct} but 10 of the 35 subjects (28%) in C_{lst} had one or more. (The reduced frequency of resuscitations in C_{lst} may be due to the reduced number of assessments necessary in the second experiment. There were 4 categories X 6 iterations = 24 estimates for each subject in experiment two as opposed to 5 categories X 9 iterations = 45 estimates for each subject in experiment one.) A Chi–square test of the differences in frequency distribution between groups yielded $\chi^2 = 11.4$, $p < .01$.

H_6

H_6 was analogous to H_3 in the fraud investigation experiment. It involved a two–way comparison of the degree of within–group agreement between C_{lst} and C_{pct} on the probability estimates of the four categories of internal control risk. Examination of histograms and descriptive statistics indicated no outliers or extreme non-normal patterns in the data matrices. However, non-parametric procedures were again deemed appropriate. Simple averages of the standard deviations of the two sets of four probabilistic internal control risk estimates were close (s_{pct} = 12.273; s_{lst} = 12.616). The average value of $|\rho|$ for each matrix of Pairwise Spearman Rank Order correlation co-efficients [$n(n-1)/2 = (4 \times 3)/2 = 6$ elements per matrix] showed slight differences, producing $|\rho|_{pct} = .4022$ and $|\rho|_{lst} = .3278$, indicating slightly higher agreement in C_{pct}. The values of Kendall's Coefficient of Concordance, which measured the overall within-group agreement on the final estimates of the probability of internal control risk categories, were both highly significant ($p < .02$).

Parametric tests were also run on the data. A one-way MANOVA compared the two sets (C_{pct} versus C_{lst}) of four probability of internal control risk estimates - first using three out of four and then four out of four estimates. The results of the tests on the vectors of the first three probability of internal control risk estimates showed that the two mean vectors were significantly different ($F = 5.692$, $p = .0016$) and Box's test for equality of covariance matrices rejected the hypothesis of equality at $p < .05$.[49] Visual examination of the profiles of the four final estimates indicated that the subjects in C_{lst} tended to be more conservative (i.e. to judge the internal control system as slightly riskier). However, that difference was not statistically significant. Figure 5. below shows the profiles of the final mean risk category estimates of the two conditions:

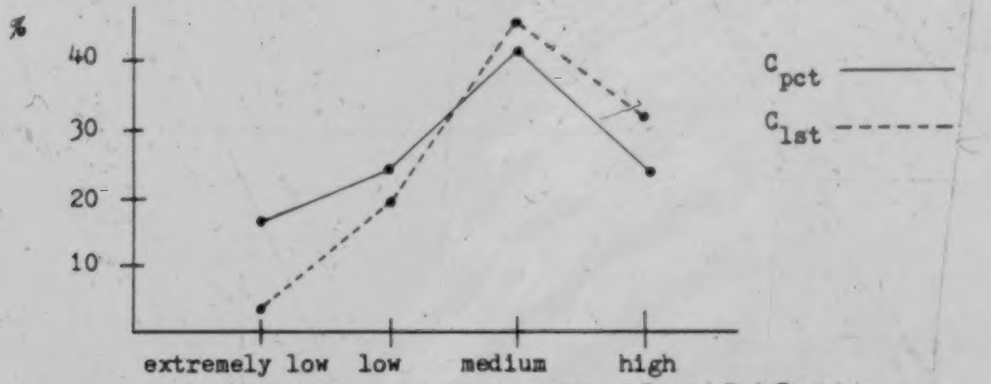


Figure 5. Mean Estimates of the Final Probability of the Internal Control Risk Categories

---

[49] No results were obtainable using the full 4 estimates as the 4 X 4 covariance matrix of estimates for C_{pct} was singular.

Overall, the results show the same kind of qualified support for H_6 as for H_3 in the first experiment. The pie chart group shows slightly higher within-group agreement as measured by average variances and intra-group correlations. However, when the high values of W_i are considered, there does not appear to be strong evidence to conclude that there is more or less agreement in either group with respect to the riskiness of the payroll Internal Control system.

H_7

H_7 concerns the variability of the subjects' judgments on the appropriate level of substantive testing. (See Appendix G.4 for summarized statistical results.) As a preliminary diagnostic, descriptive statistics and histograms were generated for the "before" and "after" scores for each of the four groups. There were no outliers and the values for skewness and kurtosis were all within acceptable parameters. A 4 X 1 ANOVA on the "before" scores of appropriate level of substantive testing did not reject the hypothesis that the four groups were equal with respect to both their mean estimates ($F = 1.652$, $p = .18$) and the variance of the estimates ($F = 1.825$, $p = .14$).

Of direct relevance to H_7 is the variablity in the "after" judgments about the the appropriate level of substantive testing. The order of the standard deviations was: $s_{lst} = .8481 < s_{nps} = 1.1006 < s_{pct} = 1.2017 < s_C = 1.3637$. A non-parametric test for the equality of variances [Conover 1980, pp. 241-42] was employed which provided for multiple comparisons if the initial results were significant. It showed the variances to be different from each other, at $p < .05$. Multiple comparisons indicated that the variance for group C was larger than for group C_{lst} at $p < .05$.

The fact that the variance of the substantive testing judgments was largest within group C suggests that an intermediate assessment of risk may be of some benefit in increasing consensus on the appropriate next step. However, although it was expected that the variance for group C would be largest, the variances for C_{pct} and C_{lst} were essentially the reverse of expectations. It appears that the processing sequence may have some effect, but the interpretation of that effect is not at all clear. These results provide some support for H_7, but that support is rather equivocal.

A somewhat unexpected result was that explicit probabilistic estimates of control risk resulted in the subjects making more conservative judgments (i.e. choosing a higher level of substantive testing) but an explicit non-probabilistic assessment of risk was associated with a less conservative judgment (a lower level of substantive testing), at least as low as no intermediate assessment. An ANOVA on the "after" scores of choice of level of substantive testing reveals a significant overall difference between means ($F = 4.607$, $p = .004$ and Newman-Keuls comparison procedure shows that the means for C and C_{nps} differ from (are higher than) the means for C_{pct} and C_{lst} at $p < .05$. These unexpected results indicate that the connection from internal control risk evaluation to choice of substantive testing appears to be more involved than expected.

H_8

H_8 predicted that the subjects in the different conditions would have differing levels of confidence in their choice of the appropriate level of substantive testing. (See Appendix G.5 for summarized statistical results.) Histograms and descriptive statistics showed that both the "before" and "after" confidence ratings to be good approximations of normally distributed variables. The hypothesis was tested by means of a 4 X 1 ANCOVA on the final confidence scores, with the initial confidence scores as a covariate.[10] The results of the ANCOVA showed a significant difference between groups ($F = 4.196$, $p < .01$). The order of the adjusted means was: $X_{pct} = 5.357 > X_{nps} = 5.051 > X_{lst} = 4.928 > X_C = 4.844$. (Box's test for the equality of variances on the "after" scores of the four groups showed that the variance in the confidence scores was equal across the four groups ($F = .833$, $p = .475$) which adds credence to the inference of differing mean confidence levels.) Paired comparisons showed a nominally significant difference between C_{pct} and C ($p < .0013$), between C_{pct} and C_{lst} ($p < .0067$), and between C_{pct} and C_{nps} ($p < .0519$). Caution must be used in interpreting these significance levels because of the multiple comparisons, but it does support the inference that the subjects in C_{pct} were more confident in their final judgments about the appropriate level of substantive testing and that the subjects in C were the least confident.

---

[10] The regression coefficient of the covarate ("before") scores was .60, significant by an F-test at $p < .001$.

Summary of H_7 and H_8

Interpretation of these results is somewhat problematic. The requirement for subjects to make explicit assessments of the risk of the internal control system prior to selecting a level of substantive testing did result in reduced variance. The results also indicated that differing processing resulting from the use of different response modes affected the level of within group agreement, but the order which resulted among the four groups did not support the superiority of the pie charts over the linear scales, contrary to expectations. It appears that if the intermediate arithmetical processing does create interference in the estimates of risk (which H_6 weakly supports), that interference does not extend to subsequent decisions based on the assessments.

However, the results did coincide with expectations completely with respect to H_8. Intermediate probabilistic estimates of internal control risk using the pie chart form of response mode were associated with higher confidence in the final judgment of the appropriate level of substantive testing. (The reader is referred to the discussion of confidence on page 49, footnote 34).

H_9

The final hypothesis, H_9, was concerned with the difficulty that the subjects in the conditions requiring explicit intermediate assessment of internal control risk (i.e. excluding group C), perceived in making those assessments. (See Appendix G.6 for summarized statistical results.) A 3 X 1 ANOVA on the difficulty scores of the subjects in the three groups yielded a highly significant result ($F = 18.088$, $p < .0001$). The mean scores were: $X_{pct} = 2.877 < X_{nps} = 3.343 < X_{lst} = 4.471$. Newman-Keuls' comparison procedure showed the means of C_{pct} and C_{nps} to be different from the mean of C_{lst} at $p < .01$, but not significantly different from each other. It is possible to make the inference that the requirement for probabilistic estimates of four different categories of internal control risk each time did not place any more cognitive demands on the subjects in C_{pct} than a single non-probabilistic estimate did on the subjects in C_{nps}. The subjects using the linear scales experienced a significantly higher level of difficulty.

Summary

The results of the Internal Control experiment strongly support the evidence from the Fraud Investigation experiment regarding the differential effects of the linear scales versus the pie charts on auditors' intuitive conformance to the patterns of the Probabilistic Judgment Paradigm. The pie chart form of response mode showed no violations of completeness and complementarity, and eliminated the resuscitations which were associated with the linear scales in conjunction with tutoring in axiomatic probability theory. Explicit evaluation of internal control risk led to greater consensus regarding the appropriate level of substantive testing and the pie chart mode of probabilistic estimation was associated with more confidence in the substantive testing decision (made based on the risk assessment) and was also perceived as no more difficult to use than the single-valued deterministic response mode.

VII. DISCUSSION

(a) Demographic Variables

The results of the tests of demographic variables indicated no significant differences by any categorization. The data was used to group the subjects by experience, firm position, educational background and prior experience with fraud investigations. However, with respect to several of the variables, there was such an uneven distribution that statistical testing was ruled out. For example, 134 (96%) of the participants had at least one university degree and only 6 (4%) did not. The same pattern obtained for prior training in probability theory – 129 (92%) of the respondents reported having had prior training in probability theory while only 11 (8%) had not. In the case of prior involvement in fraud investigation only 19 (14%) of the respondents had been involved, while 121 (86%) had not. (It might be expected that the "fraud experience" would be concentrated in the more senior members, but in fact it was very evenly distributed across the positions – $\chi^2 = 1.779$, $p = .65$). The responses of these minority respondents (i.e. no university; no probability theory; prior fraud experience) were visually scrutinized – they showed no evidence of systematic deviations from the other responses.

The Fraud Investigation Experiment

With respect to experience and position in the firm, where there was sufficient representation from the different categories, non-parametric cluster analyses (using a Euclidean distance measure) were performed on the five final probability estimates (one for each suspect) of the 35 participants in each of the four treatment conditions. Multiple runs tests [Barton and David 1957] were then done on the sequence of aggregation within each condition. There were no statistically significant patterns in any of the four treatment conditions with the runs tests done using either position in the firm or years of experience as the basis for categorization of the responses.

Also in the Fraud Investigation experiment, a number of analyses were performed to test for effects of recency of prior education. The subjects in each condition were grouped according to length of time since their university degree (which was used as a proxy for length of time since their exposure to training in probability theory). The subjects in each treatment condition were grouped into three categories – (1) those subjects two years or less from their degree, (2) those between two and five years from their degree, and (3) those more than five years from their degree. Category (3) included those subjects who did not have a university degree – each condition had one or two such subjects. (A Chi-square test of the proportions of category membership in the four treatment conditions yielded $\chi^2 = 2.118$, $p > .80$)[31] One-way MANOVA's were performed on the vectors of the subjects' mean estimates of the final probability of guilt of the suspects, categorized by "recency" as described above. Within each of the four treatment conditions there were no significant differences between any of the mean vectors.

In order to investigate the presence of any differences by recency of probability training in the process of arriving at the final estimates, a series of 20 Profile Analyses (5 suspects X 4 treatment conditions) were done over the full nine iterations. (For example, Profile Analysis #7 compared the mean profiles of the three "recency" categories in condition C_{pcnt} for suspect John Sims.) In all 20 analyses the hypothesis of parallelism was not rejected. It seems reasonable to conclude from the results that the recency of the subjects' exposure to probability training did not have any influence on their responses.

---

[31] Two results concerning the linear scale groups are relevant to this issue. First, it should be noted that none of the subjects in C_{lsnt} obeyed the axioms of completeness and complementarity in the Fraud Investigation experiment, regardless of how recent their university education. Second, the frequency of eliminations and resuscitations was very evenly distributed across the three categories in C_{lst} – $\chi^2 = .352$, $p \simeq 1.0$ in the Fraud Investigation experiment, and $\chi^2 = .146$, $p \simeq 1.0$ in the Internal Control Risk Evaluation experiment.

The Internal Control Risk Experiment

In the Internal Control Risk Evaluation experiment, analyses were also done to test for the presence of any "recency" effect. One–way MANOVA's were done on the vectors of final estimates of probabiltiy of internal control risk categories for treatment conditions C_{pct} and C_{lst}. There were no significant differences between the estimate vectors of the three "recency" categories in either condition.

As well in the Internal Control Risk experiment, two sets of four F–tests (one for each of the four treatment conditions with the responses categorized first by firm postion and then by years of experience) were done on the final choice of level of substantive testing. Within each treatment condition there were no statistically significant differences in the final choice of level of substantive testing either by position or by years of experience

The amount of time taken to complete the two experiments did not differ as a function of either position, experience, or recency of prior university training. Those participants with "medium" experience (2 – 4 years) and those who classified themselves as seniors tended to be slightly quicker, but the differences were only significant at a confidence level of $p \simeq .20$.[52]

(b) Conclusions

The evidence from these two experiments indicates that auditors' intuitive judgment processes in uncertain decision situations can be appropriately modelled by the Subjective Probability Paradigm. Using qualitative information and eliminating the need for the judges to engage in arithmetic manipulation removes the potential confound of interpreting information processing behavior resulting from the necessity of coping with the arithmetic, as violations of the principles of probability theory.

---

[52] The lack of any experience effects is consistent with the results of recent research by A. Ashton [1989] and Purvis [1989].

H_1 was confirmed. Formal tutoring in probability theory, combined with constant reminders to the judges, was effective as a means of ensuring conformance to the axioms of completeness and complementarity but it resulted in other information processing patterns, such as resuscitations, which are inconsistent with normative probability (H_2). However, the use of a simple graphical decision aid which automatically ensured completeness and complementarity without requiring any arithmetic manipulations (i.e. the pie chart), also eliminated the resuscitations.

The level of within-group agreement on the final judgments of the probability of guilt of the five suspects in the Fraud Investigation experiment was very high in C_{pct}. However, it was likewise very high in the other three groups (re the test of H_3) and it is not really possible to infer, based on this evidence, that the level of within-group agreement was higher or lower in any of the four conditions. The failure to confirm this hypothesis is not all negative. It can be interpreted as indicating that the effect of the different decision modes was neutral in that it did not significantly affect the outcome of the decision process.

The fact that the final estimates of the probability of guilt of the five suspects were not significantly different across the four treatment conditions is interesting given that both the parametric and non-parametric tests in the longitudinal analysis indicated that the process by which the participants reached those final judgments did differ across the four conditions. (Possible reasons for this convergence are discussed in the "results" section of Chapter 5.)

As an extension of the test of H_3, several analyses were performed over the full nine iterations of the experiment, in order to examine how the subjects arrived at their final probability estimates. The results of the longitudinal analysis indicated that of the four conditions, C_{lst} was the most "different" from the other three, which is consistent with the general expectation that constraining the judges to use arithmetical procedures to stay within the bounds of axiomatic probability would affect their information processing behavior.[53]

---

[53] One possible explanation for the apparent contradiction is that the cumulative effect of all the evidence was sufficiently diagnostic that all the participants came to the same final judgments, even though they used different response-mode-specific routes to get there. Alternatively, the between-group agreement could be a function of the fact

The comparison of the two probabilistic judgment conditions in the test of H_6 yielded similar results. The level of within-group agreement on the estimates of the probability of the different internal control risk categories was high in C_{pct} but it was also high in C_{lst}. It cannot be inferred that the level of agreement was different between the two.

In examining the means of the estimates, statistical tests based on the final estimates and on longitudinal analysis over the six iterations yielded consistent results. In both cases there was a significant ($p < .05$) difference between C_{pct} and C_{lst}. Participants in C_{lst} tended to judge the internal control system as riskier and also chose a slightly higher level of substantive testing.

H_7 addressed the degree of within-group agreement concerning the choice of level of substantive testing in the Internal Control experiment. One of the expectations was confirmed unequivocally. Having participants make explicit judgments about the risk of an internal control system prior to selecting an appropriate level of substantive testing did reduce the variability in that choice. However, the interpretation of the order of the variances in the three experimental conditions is not clear. The highest level of agreement (i.e. lowest variance) was achieved by participants in C_{lst}, second was C_{nps} and last was C_{pct}. The hypothesized benefits of the pie chart response mode did not materialize in this instance. It appears that intermediate judgments about internal control risk may lead to reduced variability in the final judgments about the level of substantive testing, but the form of those intermediate judgments does not seem to be important.

The experimental evidence also indicates that the approach represented by the pie chart form of response mode, which is clearly less intrusive than the separate quantitative scales, can have a beneficial effect on the outcome of the decision process. The participants employing the pie chart mode were more confident in their final judgments in both the Fraud Investigation experiment (H_4) and the Internal Control Risk Evaluation experiment (H_8). The Fraud Investigation experiment asked participants directly about their confidence in their final overall assessments of the probability of guilt of the five suspects, and the Internal Control experiment

---

[53](cont'd) that the estimates were constrained (or scaled *ex post*) to equal 100%. There was not "room" for significant variation in the final estimates – somewhat akin to the problem of restricted range in linear regression.

asked them for their confidence in their judgments of the appropriate level of substantive testing based on their estimates of the probability of the four risk levels. The results of the Fraud Investigation experiment support the inference that the pie chart form of response promoted a more integrated, cohesive view of the alternatives and that the subjects might feel that they had "a better balanced understanding" [Bowman 1968, p. 173]. The results of the Internal Control experiment indicate that choices that result from those judgments are likewise enhanced. Increased confidence in the choice of the level of substantive testing could lead to increased efficiency in the conduct of that segment of the audit.

H_5 and H_9, which dealt with the perceived difficulty of using the different response modes, (in the Fraud Investigation and Internal Control Risk Evaluation experiments respectively) were confirmed unequivocally. The results of the test of H_5 in experiment one indicate that the graphic method (pie charts) was perceived as easier than the quantitative scales, and that tutoring in probability theory to encourage adherence to the axioms of completeness and complementarity made the process even more difficult for the judges using the quantitative linear scales. This result was not surprising. If cognitive load can be inferred from scale complexity, the linear scales could be expected to impose considerably greater cognitive demands on the participants. The results of the statistical tests confirm that inference.

The results of the test of H_9 in experiment two were based on the difficulty ratings of three different conditions - two repeated from experiment one, and a non-probabilistic mode requiring only a single estimate each iteration. The fact that the mean rating of the judges in C_{pct} was significantly less than the mean rating of judges in C_{lst} confirms the results of H_5, but in addition, the fact that C_{pct} was perceived to be at least as easy as C_{nps}, which required only one non-probabilistic estimate each time, indicates the degree of benefit in terms of reduced cognitive load.[54]

---

[54] An interesting alternate perspective is presented in recent research by Hamm [1985] and Hammond, Hamm, Grassia, and Pearson [1987]. They distinguish between analytical judgment and intuitive judgment. The authors suggest that the nature of the judgment process (analytical vs. intuitive) which a decision maker uses, is largely determined by the nature of the task. They propose a "task continuum index" (TCI) which locates

The emphasis on a more realistic experimental setting was important. In order for the results of the experiments to allow any inferences about auditors' information processing behavior in a "typical" uncertain decision situation, the experiment must at least contain the characteristics of that situation. A series of quantitative probability statements to be combined according to some mathematical rule (such as Bayes' Theorem) is definitely not characteristic of such a situation. Even if all the judges were able to compute a quantitative answer in perfect agreement with Bayes' Theorem, that hardly speaks to the question of how auditors make judgments in their regular professional practice.[55] Professional judgment involves interpreting and integrating qualitative cues to arrive at a judgment. The evidence from these experiments is that in such situations the subjective probability model is a valid descriptive model.

(c) Limitations and Suggestions for Future Research

The overall goal of the experiments was to support on a qualitative level, the validity of the probabilistic judgment paradigm as a descriptive model of auditors' decision making under uncertainty, and to test the value of a simple graphical decision aid to be used in such decision situations.

---

[54](cont'd) tasks as intuition-inducing or analysis-inducing. According to their criteria, the present experimental tasks would be classified towards the intuition-inducing end of the continuum (e.g. the evidence is indirect and qualititative in nature). The two forms of processing employed in the experiment can be considered as being either congruent with these intuition-inducing tasks (the visually-based response mode) or incongruent with these tasks (the quantitive scales). Thus the results of the experiments can be interpreted as providing positive evidence on the role of probability in people's intuitive judgments, and they can also be interpreted as indirectly supporting the experimental results of Hammond et al. [1987] which indicated that the quality of judgment is an increasing function of the congruity between the nature of the task (as measured by the TCI) and the form of cognitive activity required of the judge.

[55] The emphasis on a realistic setting was also intended to make the experimental task more "involving" and thereby more motivating for the auditors. An increased degree of interest and committment to the task could be expected to enhance the validity of the responses. Informal comments and discussion with subjects after the sessions indicated a high level of involvement.

One aspect of this research paradigm which has been suggested as a potential limitation in the context of research into such an applied field as auditing is the very fact that it is an experiment. Of necessity it abstracts from a "real-life" setting and presents subjects with a very abridged scenario. Although considerable effort was made to minimize the potential negative effects of the experimental paradigm (e.g. the experiments were conducted in the offices of the participants; considerable care was taken to develop realistic case descriptions for the experiments), the fact remains that this research followed the experimental paradigm. The increased internal validity afforded by this methodology was felt to outweigh any potential reduction in external validity. This reflects the personal views of the researcher, but it should also be noted that the experimental paradigm is widely accepted as an appropriate methodology for conducting investigations into the nature of probabilistic judgment in both cognitive psychology and behavioral accounting research.

The results of these experiments suggest several avenues for future research. One limitation of these studies which could be addressed by future research relates to the emphasis on the use of qualitative indirect evidence. The "strength" of using indirect qualitative evidence to increase validity creates a countervailing weakness - there are no objective standards against which to compare subjects' responses. The use of "soft" measures such as confidence and consensus is not as good. This is a problem which has plagued behavioral auditing research almost since its inception. It is difficult to compare the benefits of one method versus another without some sort of objective standard by which to measure the results. In order to properly resolve the issue of the appropriateness of the different models and approaches, it will be necessary to create an experiment which will avoid the confound of an intervening variable such as the need for arithmetical processing, while allowing the results to be measured against concrete objective standards. The accomplishment of this difficult task will be a significant contribution.

Another extension of this research could be to address the issue of whether or not ther is any relation between choice of processing mode and judgment situation in audit decision making. These experiments have shown that auditors can process information in a probabilistic mode and that their intuitive patterns are consistent with axiomatic probability, but the experiments did not address the interesting question of what is the preferred mode of processing. If subjects were given a number of different alternative processing modes, probabilistic and non-probabilistic, which one would they choose? It may be that the choice of preferred mode would be different for different types of uncertain judgments, such as routine versus non-routine. An experiment designed to allow the auditor-subjects to choose the response mode from several probabilistic and non-probabilistic alternatives in different judgment situations could be employed to answer this question.

An experimental method which could shed considerable light on what the subjects were doing in these instances is protocol analysis. If guided by the theoretical framework of the Probabilistic Judgment Paradigm, such an experimental methodology could be employed to perform a deeper analysis of the subjects' information processing behavior during this type of decision. Protocol analysis would also be an appropriate way to investigate the question of preference for processing mode, as well as ability to use it.

An area of futher investigation is also suggested by the observation that the one property of Bayesian revision that was violated by all subjects related to the re-allocation of the remaining probabilities to the other alternatives when the probability of one alternative was changed. For instance, in the Fraud Investigation experiment, if the probability of guilt of Arthur Denton was increased by 20% as a result of the fourth piece of evidence, in order to comply with the correct pattern of Bayesian revision, the proper procedure would be to reduce each of the other four suspects by an equal amount - 5%. None of the participants in any of the conditions did this. While the normatively correct reassignment of probabilities in such cases can be quite arithmetically complex and is certainly non-intuitive, the subjects' behavior could also be interpreted as evidence of what could be termed "contingent configurality". Often a subject added the "extra" probability or subtracted the reduced probability to/from one or perhaps two suspects. It appears that the subjects were considering the suspects/evidence in an interdependent fashion, even though each piece of evidence satisfied conditional independence - i.e. it referenced one and only one suspect at a time. The methodology of protocol analysis could be employed to gain insight into what cognitive procedures or "rules" the subjects were using to distribute a change in the probability of one alternative to the other alternatives.

Summary

Overall, it seems reasonable to conclude that the results of these experiments support the validity of the Probabilistic Judgment Paradigm as the conceptual underpinning of auditors' decision making under uncertainty. A methodology which allows the judges to use the Probabilistic Judgment Paradigm without getting "mired" in the arithmetic enhances the secondary characteristics of the judgments, including the degree of within-group consensus, the speed with which the judges can make probabilistic assessments and the confidence they have in those assessments (and in subsequent judgments based on those assessments).

REFERENCES

Abdel-khalik, A.R., D. Snowball, and J.H. Wragge, "The Effects of Certain Internal Audit Variables on the Planning of the External Audit," Accounting Review (April 1983), pp. 215-27.

Abdolmohammadi, M.J., "Efficiency of the Bayesian Approach to Compliance Testing: Some Empirical Evidence," Auditing: A Journal of Practice and Theory (Spring 1986), pp. 1-16.

Abdolmohammadi, M.J., and P.D. Berger, "A Test of the Accuracy of Probability Assessment Techniques in Auditing," Contemporary Accounting Research (Fall 1986), pp. 149-65.

Agnew, N.H., R.A. Agnew, J. Rasmussen, and K.R. Smith, "An Application of Chance Constrained Programming to Portfolio Selection in a Casualty Insurance Firm," Management Science (June 1969), pp. 512-20.

Altman, E.I., "Multi-Dimensional Graphics and Bankruptcy Prediction: A Comment," Journal of Accounting Research (Spring 1983), pp. 279-99.

Altman, E.I., "Accounting Implications of Failure Prediction," Journal of Accounting, Auditing and Finance (Fall 1982), pp. 4-19.

Altman, E.I., "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," Journal of Finance (September 1968), pp. 589-609.

Anderson, RJ, The ExternalAudit (2nd edition), Copp Clark Pittman: Toronto, 1984.

Ashton, A.H., "Experience Effects in Auditing: The Role of Task-Specific Knowledge," Paper presented at the 1989 Audit Judgment Symposium, Center for Accounting Research, University of Southern California, Los Angeles, California, February 20-21, 1989.

Ashton, A.H., "Does Consensus Imply Accuracy in Accounting Studies of Decision Making?" Accounting Review (April 1985), pp. 173-86.

Ashton, A.H., and R.H. Ashton, "Sequential Belief Revision by Auditors: Tests of a New Model," Accounting Review (October 1988), pp. 623-41.

Ashton, R.H., Research in Audit Decision Making: Rationale, Evidence, and Implications, Research Monograph No. 6, Vancouver, B.C.: The Canadian Certified General Accountants' Research Foundation, 1983.

Ashton, R.H., "Human Information Processing Research in Auditing: A Review and Synthesis," in Nicholls, D.R., and H.F. Stettler (eds.), Auditing Symposium VI, Lawrence, Kansas: University of Kansas, 1982, pp. 71-83 (a).

Ashton, R.H., Human Information Processing in Accounting, Studies in Accounting Research No. 17, Sarasota, Florida: AAA, 1982 (b).

Ashton, R.H., "An Experimental Study of Internal Control Judgments," Journal of Accounting Research (Spring 1974), pp. 143-57 (a).

Ashton, R.H., "Cue Utilization and Expert Judgments: A Comparison of Independent Auditors with Other Judges," Journal of Applied Psychology (August 1974), pp. 437–44 (b).

Ashton, R.H., and P.R. Brown, "Descriptive Modelling of Auditors' Internal Control Judgments: Replication and Extension," Journal of Accounting Research (Spring 1980), pp. 269–77.

Ashton, R.H., and S.S. Kramer, "Students as Surrogates in Behavioral Accounting Research: Some Evidence," Journal of Accounting Research (Spring 1980), pp. 1–15.

Atkinson, A.B., "On the Measurement of Inequality," Journal of Economic Theory (September 1970), pp. 244–63.

Atkinson, J.W., "Motivational Determinants of Risk Taking Behavior," Psychological Review (1957), pp. 359–72.

Baiman, S., "Agency Theory Research in Managerial Accounting: A Survey," Journal of Accounting Literature (1982), pp. 154–213.

Barton, D.E., and F.N. David, "Multiple Runs," Biometrika (1957), 54, pp. 168–78.

Baumol, W.J., "An Expected Gain – Confidence Limit Criterion for Portfolio Selection," Management Science (October 1963), pp. 174–82.

Beach, L.R., "Probability Magnitudes and Conservative Revision of Subjective Probabilities," Journal of Experimental Psychology (May 1968), pp. 57–63.

Beck, P.J., I. Solomon, and L.A. Tomassini, "Subjective Prior Probability Distributions and Audit Risk," Journal of Accounting Research (Spring 1985), pp. 37–56.

Becker, S.W., and F.O. Brownson, "What Price Ambiguity? Or the Role of Ambiguity in Decision Making," Journal of Political Economy (February 1964), pp. 62–73.

Bernoulli, D., *Exposition of a New Theory on the Measurement of Risk*, Farnsborough Hants, England: Gregg Press, 1967 (original work published 1738).

Biggs, S.F., W.F. Messier Jr., and J.V. Hansen, "A Descriptive Analysis of Computer Audit Specialists' Decision Making Behaviour in Advanced Computer Environments," Auditing: A Journal of Practice and Theory (Fall 1987), pp. 1–22.

Birnberg, J.C., and J.F. Shields, "Three Decades of Behavioral Accounting Research: A Search for Order," Behavioral Research in Accounting (1989), 1, pp. 23–74.

Bock, R.D., *Multivariate Statistical Methods in Behavioral Research*, New York: McGraw Hill, 1975.

Boneau, C.A., "The Effects of Violations of Assumptions Underlying the t–Test", Psychological Bulletin (January 1960), pp. 49–64.

Bowman, W.J., *Graphic Communication*, New York: John Wiley and Sons, 1968.

Box, G.E.P., "Non–Normality and Tests on Variances", Biometrika (December 1953), pp. 318–35.

Brown, H.B., and A.B. Forsythe, "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means," Technometrics (1974), 16, pp. 129–32. (a)

Brown, H.B., and A.B. Forsythe, "Robust Tests for the Equality of Variances," Journal of the American Statistical Association (1974), 69, pp. 364-67. (b)

Brunswik, E., Perception and the Representative Design of Psychological Experiments, Los Angeles, Ca.: Unversity of California, 1956.

Brunswik, E., The Conceptual Framework of Psychology, Chicago, Ill.: University of Chicago, 1952.

Budescu, D.V., and T.S. Wallsten, "Subjective Interpretation of Precise and Vague Uncertainties," in Wright, G., and P. Ayton (eds.), Judgmental Forecasting, New York: John Wiley and Sons, 1987, pp. 63-82.

Budescu, D.V., and T.S. Wallsten, "Dyadic Decisions With Numerical and Verbal Probabilities," Working Paper, University of North Carolina, Chapel Hill, N.C., 1988.

Burgstahler, D., and G.L. Sundem, "The Evolution of Behavioral Accounting Research in the United States, 1967-1987," Behavioral Research in Accounting (1989), 1, pp. 75-108.

Campbell, D.T., and J.C. Stanley, Experimental and Quasi-Experimental Designs for Research, Chicago, Ill.: Rand McNally, 1966.

Chernoff, H., "The Use of Faces to Represent Points in K-Dimensional Space Graphically," Journal of the American Statistical Association (1973), 68, pp. 361-68.

Chernoff, H., "Graphical Representation as a Discipline", in Wang, P.C.C. (ed.), Graphical Representation of Multivariate Data, New York: Academic Press, 1978.

Chesley, G.R., "Inference in Auditing: The Implications of Coherence," Proceedings of the 5th Conference on Contemporary Accounting Research, McMaster University, 1984, pp. 97-107.

Chesley, G.R., "Procedures for the Communication of Uncertainty in Auditors' Working Papers," in Burns, T.H. (ed.), Behavioral Experiments in Accounting II, Columbus, Ohio: Ohio State University, 1979, pp. 115-49.

Chesley, G.R., "Subjective Probability Elicitation Techniques: A Performance Comparison," Journal of Accounting Research (Autumn 1978), pp. 225-41.

Chesley, G.R., "Subjective Probability Elicitation: The Effect of Congruity of Datum and Response Mode on Performance," Journal of Accounting Research (Spring 1977), pp. 1-11.

Chesley, G.R., "The Elicitation of Subjective Probabilities: A Laboratory Study in an Accounting Context," Journal of Accounting Research (Spring 1976), pp. 27-48.

CICA Handbook, Toronto: CICA, 1987.

Cohen, L.J., The Probable and the Provable, Oxford, England: Oxford University Press, 1977.

Cooley, W.W., and P.R. Lohnes, Multivariate Data Analysis, New York: John Wiley and Sons, 1971.

Coombs, C.H., Portfolio Theory: A New Theory of Risky Decision Making, La Decision, Paris: Centre National de la Recherche Scientifique, 1969.

Coombs, C.H., and J.N. Bowen, "A Test of VE-Theories of Risk and The Effect of the Central Limit Theorem," Acta Psychologica (January 1971), pp. 15-28.

Coombs, C.H., and D.G. Pruitt, "Components of Risk in Decision Making: Probability and Variance Preferences," Journal of Experimental Psychology (November 1960), pp. 265–77.

Corless, J.C., "Assessing Prior Distributions for Applying Bayesian Statistics in Auditing," Accounting Review (July 1972), pp. 556–66.

Crosby, M.A., "Bayesian Statistics in Auditing: A Comparison of Probability Elicitation Techniques," Accounting Review (April 1981), pp. 355–65.

Crosby, M.A., "Implications of Prior Probability Elicitation on Auditor Sample Size Decisions," Journal of Accounting Research (Autumn 1980), pp. 585–93.

Croxton, F.E., and R.E. Stryker, "Bar Charts Versus Circle Diagrams," Journal of the American Statistical Association (1927), 22, pp. 473–82.

Croxton, F.E. and H. Stein, "Graphic Comparisons by Bars Squares, Circles and Cubes," Journal of the American Statistical Association (1932), 27, pp. 54–60.

Davidson, R.A., Behavioral Research in Auditing," Working Paper, University of Calgary, June, 1989.

Dawes, R., and B. Corrigan, "Linear Models in Decision Making," Psychological Bulletin (February 1974), pp. 95–106.

Demski, J.S., Information Analysis (2nd edition), Reading, Mass.: Addison-Wesley, 1980.

Dickhaut, J.W., "Alternative Information Structures and Probability Revisions," Accounting Review (January 1973), pp. 61–79.

Dixon, W.J. (ed.). BMDP Statistical Software: 1983 Printing with Additions, Los Angeles, California: University of California Press, 1983.

Donaldson, T.S., "Robus[illegible] the F-Test to Errors of Both Kinds and the Correlation Between the Numerator and the Denominator of the F-Ratio," Journal of the American Statistical Association (June 1968), pp. 660–76.

DuCharme, W.M., and M.L. Donnell, "Intrasubject Comparison of Four Response Modes for "Subjective Probability" Assessment," Organizational Behavior and Human Performance (August 1973), pp. 108–17.

Edwards, A.L., Experimental Design in Psychological Research (5th edition), New York: Harper and Row, 1985.

Edwards, A.L., Multiple Regression and the Analysis of Variance and Co-Variance, San Francisco: W.H. Freeman and Company (1979).

Edwards, A.L., Probability and Statistics, New York: Holt, Rinehart and Winston, 1971.

Edwards, W., "Conservatism in Human Information Processing," in Kleinmuntz, B. (ed.), Formal Representation of Human Judgment, New York: John Wiley and Sons, 1968, pp. 17–52.

Edwards, W., "The Prediction of Decisions Among Bets," Journal of Experimental Psychology (September 1955), pp. 201–214.

Edwards, W., "The Theory of Decision Making," Psychological Bulletin (July 1954), pp. 380–417.

Edwards, W., H. Lindeman, and L.D. Phillips, "Emerging Technologies for Making Decisions," in New Directions in Psychology II, New York: Holt, Rinehart and Winston, 1965, pp. 261-325.

Edwards, W., H. Lindeman, and L.J. Savage, "Bayesian Statistical Inference for Psychological Research," Psychological Review (May 1963), pp. 193-242.

Eells, W.C., "The Relative Merits of Circles and Bars for Representing Component Parts," Journal of the American Statistical Association (1926), 21, pp. 364-77.

Einhorn, H., "Expert Judgment: Some Necessary Conditions and an Example," Journal of Applied Psychology (October 1974), pp. 562-71.

Einhorn, H., and R. Hogarth, "Ambiguity and Uncertainty in Probabilistic Inference," Psychological Bulletin (1985), pp. 433-61.

Einhorn, H., and R. Hogarth, "Behavioral Decision Theory: Processes of Judgment and Choice," Annual Review of Psychology (1981), pp. 53-88.

Ellsburg, D., "Risk, Ambiguity, and the Savage Axioms," Quarterly Journal of Economics (November 1961), pp. 643-69.

Emby, C., "Models of Risky Decision Behavior," unpublished manuscript, University of Alberta, 1985.

Emby, C., and M. Gibbins, "Good Judgment in Public Accounting: Quality and Justification," Contemporary Accounting Research (Spring 1988), pp. 287-313.

Emory, W., Business Research Methods (revised edition), Homewood, Ill.:Irwin, 1980.

Farquhar, A.B., and H. Farquhar, Economics and Industrial Delusions: A Discourse on the Case for Protection, New York: Putnam and Sons, 1891.

Felix, W.L., "Evidence on Alternative Means of Assessing Prior Probability Distributions for Audit Decision Making," Accounting Review (October 1976), pp. 800-07.

Felix, W.L., and M.S. Niles, "Research in Internal Control Evaluation," Auditing: A Journal of Practice and Theory (Spring 1988), pp. 43-60.

Fhaner, S., "Subjective Probability and Everyday Life," Scandinavian Journal of Psychology (1st Quarter 1977), pp. 81-4.

Firth, M., "Consensus Views and Judgmental Models in Materiality Decisions," Accounting, Organizations and Society (1979), 4, pp. 283-295.

Fischoff, B., and R. Beyth-Marom, "Hypothesis Evaluation From a Bayesian Perspective," Psychological Review (July 1983), pp. 239-60.

Fischoff, B., P. Slovic, and S. Lichtenstein, "Knowing With Certainty: The Inappropriateness of Extreme Confidence," Journal of Experimental Psychology: Human Perception and Processes (1977), pp. 352-64.

Fishburn, P.C., "Mean-Risk Analysis with Risk Associated With Below-Target Returns," American Economic Review (March 1977), pp. 116-26.

Fishburn, P.C., "The Axioms of Subjective Probability," Statistical Science (1986),1, pp. 335-58.

Fox, D.J., and K.E. Guire, Documentation for MIDAS, Ann Arbor, Michigan: The Statistical Research Laboratory, The University of Michigan, 1976.

Gabrielsson, A., and P. Seeger, "Tests of Significance in Two-Way Designs (Mixed Model) With Dichotomous Data," British Journal of Mathematical Statistics (1971), 24, pp. 111-16.

Gaumnitz, B.R., T.R. Nunamaker, J.J. Surdick, and M.F. Thomas, "Auditor Consensus in Internal Control Evaluation and Audit Program Planning," Journal of Accounting Research (Autumn 1982, Pt. II), pp. 745-55.

Gibbins, M., "Human Inference, Heuristics, and Auditors' Judgment Process," CICA Audit Research Symposium, 1977.

Gibbins, M., "Propositions About the Psychology of Professional Judgment in Public Accounting," Journal of Accounting Research (Spring 1984), pp. 103-25.

Gibbs, T.E., and R.G. Schroeder, "Evaluating the Competence of Internal Audit Departments," in Symposium on Auditing Research II, Urbana, Ill.: University of Illinois, 1979, pp. 207-25.

Gnanadesikan, R., "Graphic Data Analysis: Issues, Tools, and Examples," Paper presented at the Annual Meeting of the American Association of Advertising Science, San Francisco, California, 1980.

Greenhouse, S., and S. Geisser, "On Methods in the Analysis of Profile Data," Psychometrika (1959), 24, pp. 95-112.

Hamilton, R.E., and R.F. Wright, "Internal Control Judgments and Effects of Experience: Replications and Extensions," Journal of Accounting Research (Autumn 1982, Pt. II), pp. 756-65.

Hamm, R.H., "Moment by Moment Variation in the Cognitive Activity of Experts," Report No. 257, Center for Research on Judgment and Policy, University of Colorado, Boulder, Colorado, August, 1985.

Hammond, K.R., R.H. Hamm, J. Grassia, and T. Pearson, "Direct Comparison of the Efficacy of Intuitive and Analytical Cognition in Expert Judgment," IEEE Transactions on Systems, Man, and Cybernetics (September/October 1987), pp. 753-70.

Haskins, M.E., "Client Control Environments: An Examination of Auditors' Perceptions," Accounting Review (July 1987), pp. 542-63.

Hays, W.L., Statistics (3rd edition), New York: CBS College Publishing, 1981.

Hofstedt, T.R., and G.D. Hughes, "An Experimental Study of the Judgment Element in Disclosure Decisions," Accounting Review (April 1977), pp. 379-95.

Hunyh, H., and L.S. Feldt, "Conditions Under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions," Journal of the American Statistical Association (1970), 65, pp. 1582-89.

Huynh, H., and L.S. Feldt, "Performance of Traditional F-Tests in Repeated Measures Designs Under Covariance Heterogeneity," Communications in Statistics - Theoretical Methods (1980), 1, pp. 61-74.

Ives, B., "Graphical User Interfaces for Business Information Systems," Management Information Systems Quarterly (Special Issue 1982), pp. 15-47.

Johnson, W.B., "Representativeness in Judgmental Predictions of Corporate Bankruptcy," Accounting Review (January 1983), pp. 78–97.

Johnson, R.A., and D.W. Wichern, Applied Multivariate Statistical Analysis, Englewood Cliffs, N.J.: Prentice-Hall, 1982.

Joyce, E.J., "Expert Judgment in Audit Program Planning," in Studies on Human Information Processing in Accounting, supplement to Journal of Accounting Research (1976), pp. 29–60.

Joyce, E.J., and G. Biddle, "Anchoring and Adjustment in Probabilistic Inference in Auditing," Journal of Accounting Research (Spring 1981), pp. 120–45 (a).

Joyce, E.J., and G. Biddle, "Are Auditors' Judgments Sufficiently Regressive?" Journal of Accounting Research (Autumn 1981), pp. 323–49 (b).

Joyce, E.J., and R. Libby, "Behavioral Studies of Audit Decision Making," Journal of Accounting Literature (1982), pp. 103–23.

Kahneman, D., and A. Tversky, "Prospect Theory: An Analysis of Decisions Under Risk," Econometrica (March 1979), pp. 263–91 (a).

Kahneman, D., and A. Tversky, "On The Interpretation of Intuitive Probability: A Reply to Jonathan Cohen," Cognition (December 1979), pp. 409–11 (b).

Kahneman, D., P. Slovic, and A. Tversky, Judgment Under Uncertainty: Heuristics and Biases, New York: Cambridge University Press, 1982.

Kennedy, H.A., "A Behavioral Study of the Usefulness of Four Financial Ratios," Journal of Accounting Research (Spring 1975), pp. 97–116

Kida, T., "The Effects of Causality and Specificity on Data Use," Journal of Accounting Research (Spring 1984), pp. 145–52.

Kinney, W.R. Jr., "A Decision Theory Approach to the Sampling Problem in Auditing," Journal of Accounting Research (Spring 1975), pp. 117–32.

Kinney, W.R. Jr., and W.C. Uecker, "Mitigating the Consequences of Anchoring in Auditor Judgments," Accounting Review (January 1982), pp. 55–69.

Knight, F.H., Risk, Uncertainty, and Profit, New York: Kelley and Millman, 1957.

Krickeberg, K., Probability Theory, Reading, Mass.: Addison-Wesley, 1965.

Larssen, S., and G.R. Chesley, "An Analysis of the Auditor's Uncertainty About Probabilities," Contemporary Accounting Research (Fall 1986), pp. 259–82.

Lawrence, W.E., Theory of Probability, Dubuque, Iowa: W.C. Brown, 1967.

Leslie, D.A., A.D. Teitlebaum, and R.J. Anderson, Dollar Unit Sampling: A Practical Guide for Auditors, London: Pittman Publishing Ltd., 1980.

Lewis, N. (ed.), The New Roget's Thesaurus in Dictionary Form, New York: Putnam's 1964.

Libby, R., Accounting and Human Information Processing: Theory and Applications, Englewood Cliffs, N.J.: Prentice-Hall Inc., 1981.

Libby, R., "Bankers' and Auditors' Perceptions of the Message Communicated by the Audit Report," Journal of Accounting Research (Spring 1979), pp. 99–122 (a).

Libby, R., "The Impact of Uncertainty Reporting on the Loan Decision," Supplement to Journal of Accounting Research (1979), pp. 37–57 (b).

Libby, R., and P.C. Fishburn, "Behavioral Models of Risk-Taking in Business Decisions: A Survey and Evaluation," Journal of Accounting Research (Autumn 1977), pp. 272–92.

Lichtenstein, S., "Bases for Preferences Among Three-Outcome Bets," Journal of Experimental Psychology (February 1965), pp. 162–69.

Lichtenstein, S., and L.D. Phillips, "Calibration of Probabilities: The State of the Art," in Jungermann, H., and G. deZeuuw (eds.), Decision Making and Change in Human Affairs , Dordrecht, Holland: Reidel, 1977, pp. 275–324.

Lopes, L., "Risk and Distributional Inequality," Journal of Experimental Psychology: Human Perception and Processes (August 1984), pp. 465–85.

Lopes, L., "Between Hope and Fear: The Psychology of Risk," Working Paper, University of Wisconsin-Madison, 1986.

MacCrimmon, K.R., "Descriptive and Normative Implications of the Decision Theory Postulates," in K. Borsch and J. Mossin (eds.), Risk and Uncertainty, New York: MacMillan, 1968, pp. 3–32.

McClelland, D.C., The Achieving Society, Princeton, N.J.: Van Nostrand, 1961.

Markowitz, H., "Portfolio Selection," Journal of Finance (March 1952), pp. 77–91.

Markowitz, H., Portfolio Selection: Efficient Diversification of Investments, New York: John Wiley and Sons, 1959.

Mock, T., and J.L. Turner, "The Effect of Changes in Internal Controls on Audit Programs," in Burns, T.J. (ed.), Behavioral Experiments in Accounting II, Columbus: Ohio Ohio State University, 1979, pp. 277–302.

Mock, T., and I. Vertinsky, Risk Assessment in Accounting and Auditing: a Research Report, Research Monograph No. 10, Vancouver, B.C.: The Canadian Certified General Accountants' Research Foundation, 1985.

Mood, A.M., "The Distribution Theory of Runs," Annals of Mathematical Statistics (1940), 11, pp. 367–92.

Moriarity, S., "Communicating Financial Information Through Multi-Dimensional Graphics," Journal of Accounting Research (Spring 1979), pp. 205–24.

Moriarity, S., and F.H. Barron, "Modeling the Materiality Judgments of Partners," Journal of Accounting Research (Autumn 1986), pp. 320 41.

Moriarity, S., and F.H. Barron, "Judgment Based Definition of Materiality," Supplement to Journal of Accounting Research (1979), pp. 114–35.

Myers, J.L., Fundamentals of Experimental Design (3rd ed.), Boston: Allyn and Bacon, 1979.

Neter, J., and J.K. Loebbeke, Behavior of Major Statistical Estimators in Sampling Accounting Populations , Audit Research Monograph No. 2. Sarasota, Florida: AAA, 1975.

Newell, A., and H.A. Simon, Human Problem Solving, Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Payne, J.W., and M.L. Braunstein, "Preferences Among Gambles With Equal Underlying Distributions," Journal of Experimental Psychology (January 1971), pp. 13-18.

Peterson, C.R., "Aggregating Information About Signal and Noise," in Proceedings of the 76th Annual Convention of the American Psychological Association, 1968, pp. 123-24.

Phillips, L., "The "True" Probability Problem," Acta Psychologica (1970), pp. 254-64.

Pitz, G.F., and N.J. Sachs, "Judgment and Decision: Theory and Applications," Annual Review of Psychology (1984), pp. 139-63.

Playfair, W., The Commercial and Political Atlas, London: Corry, 1786.

Pollatsek, I., and A. Tversky, "A Theory of Risk," Journal of Mathematical Psychology (December 1970), pp. 540-53.

Purvis, S.E.C., "Audit Expertise: The Relation Between Error Frequency Knowledge and Experience," Paper presented at the 1989 Audit Judgment Symposium, Center for Accounting Research, University of Southern California, Los Angeles, California, February 20-21, 1989.

Raiffa, H., Decision Analysis, Reading, Mass: Addison-Wesley, 1968.

Rapoport, A., and T.S. Wallsten, "Individual Decision Behavior," Annual Review of Psychology(1972), pp. 131-76.

Reckers, P.M.J., and M.E. Taylor, "Consistency in Auditors' Evaluations of Internal Accounting Controls," Journal of Accounting, Auditing and Finance (Fall 1979), pp. 42-55.

Remus, W., "A Study of Graphical and Tabular Displays and Their Interaction with Environmental Complexity," Management Science (September 1987), pp. 1200-04.

Report of the Commission to Study the Pulic's Expectation of Audits (MacDonald Commission Report), Toronto: CICA, 1988.

Robinson, L.B., and R. Hastie, "Revision of Beliefs When a Hypothesis is Eliminated from Consideration," Journal of Experimental Psychology: Human Perception and Performance (August 1985), pp. 443-56.

Savage, L.J., The Foundations of Statistics, Homewood, Ill.: John Wiley and Sons, 1954.

Schaefer, R.E., and K. Borcherding, "The Assessment of Subjective Probability Distributions: A Training Experiment," Acta Psychologica (April 1973), pp. 117-29.

Schum, D.A., and A.W. Martin, "Formal and Empirical Research on Cascaded Inference in Jurisprudence," Law and Society Review (1982), pp. 105-51.

Schneider, S.L., and L. Lopes, "Reflection in Preferences Under Risk: Who and When May Suggest Why," Journal of Experimental Psychology: Human Perception and Processes (November 1986), pp. 535-48.

Scott, W.R., "The State of the Art in Academic Research in Auditing," Journal of Accounting Literature (1984), pp. 153-200.

Seber, G.A.F., Multivariate Observations, New York: John Wiley and Sons, 1984.

Shafer, G., A Mathematical Theory of Evidence, Princeton, N.J.: Princeton University Press, 1976.

Shafer, G., and R. Srivasta, "The Bayesian and Belief Function Formalisms I: A General Perspective for Auditing," Working Paper, School of Business, University of Kansas, February 1987.

Shafer, G., and A. Tversky, "Language and Designs for Probability Judgment," Cognitive Science (July/September 1985), pp. 309-39.

Shanteau, J., "The Role of Communication, Confidence, and Creativity for Expertise," Paper presented at the 1989 Audit Judgment Symposuim, Center for Accounting Research, University of Southern California, Los Angeles, California, February 20-21, 1989.

Slovic, P., "The Relative Influence of Probabilities and Payoffs Upon Perceived Risk of a Gamble," Psychonomic Science (October 1967), pp. 223-24.

Slovic, P., and S. Lichtenstein, "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment," Organizational Behavior and Human Performance (November 1971), pp. 649-744.

Slovic, P., and S. Lichtenstein, "Relative Importance of Probabilities and Payoffs in Risk Taking," Journal of Experimental Psychology Monograph (1968) (a).

Slovic, P., and S. Lichtenstein, "Importance of Variance Preferences in Gambling Decisions," Journal of Experimental Psychology (December 1968), pp. 646-54 (b).

Slovic, P., B. Fischoff, and S. Lichtenstein, "Behavioral Decision Theory," Annual Review of Psychology (1977), pp. 1-39.

Solomon, I., "Multi-auditor Judgment/Decision Making Research," Journal of Accounting Literature (1987), 6, pp. 1-25.

Solomon, I., "Probability Assessment by Individual Auditors and Audit Teams: An Empirical Investigation," Journal of Accounting Research (Autumn 1982, Pt. II), pp. 689-710.

Solomon, I., J.L. Krogstad, M.B. Romney, and L.A. Tomassini, "Auditors' Prior Probability Distributions for Account Balances," Accounting, Organizations and Society (1982), 7, pp. 27-41.

Sorenson, J.E., "Bayesian Analysis in Auditing," Accounting Review (July 1969), pp. 555-61.

Stael von Holstein, C.-A.S., "Measurement of Subjective Probability," Acta Psychologica (December 1970), pp. 146-59.

Stael von Holstein, C.-A.S., "Probabilistic Forecasting: An Experiment Related to the Stock Market," Organizational Behavior and Human Performance (August 1972), pp. 139-58.

Statement of Auditing Standards No. 3, New York: AICPA, 1963.

Statement of Auditing Standards No. 53, New York: AICPA, 1988.

Stock, D., and C.J. Watson, "Human Judgment Accuracy, Multidimensional Graphics, and Human Versus Models," Journal of Accounting Research (Spring 1984), pp. 192-206.

Stoloff, P.H., "Correcting for Heterogeneity of Covariance for Repeated Measures Designs of the Analysis of Variance," Educational and Psychological Measurement (1970), 30, pp. 909–24.

Swieringa, R.J., M. Gibbins, J. Sweeney, and J. Larson, "Experiments in the Heuristics of Information Processing," in Studies on Human Information Processing in Accounting, Supplement to Journal of Accounting Research (1976), pp. 159–87.

Teigen, K.H., "Studies in Subjective Probabilities III: The Unimportance of Alternatives," Scandinavian Journal of Psychology (2nd Quarter 1983), pp. 97–105.

Teigen, K.H., "Subjective Sampling Distributions and the Additivity of Estimates," Scandinavian Journal of Psychology (1st Quarter 1974), pp. 50–55 (a).

Teigen, K.H., "Overestimation of Subjective Probabilities," Scandinavian Journal of Psychology (1st Quarter 1974), pp. 56–62 (b).

Thorndyke–Barnhardt Comprehensive Desk Dictionary, Garden City, N.Y.: The Country Life Press, 1965.

Tiku, M.L., "Power Function of the F–Test Under Non–Normal Situations," Journal of the American Statistical Association (December 1971), pp. 913–916.

Tomassini, L.A., I. Solomon, M.B. Romney, and J.L. Krogstad, "Calibration of Auditors' Probabilistic Judgments: Some Evidence," Organizational Behavior and Human Performance (December 1982), pp. 391–406.

Torgerson, W.S., Theory and Methods of Scaling, New York: John Wiley and Sons, 1958.

Tversky, A., "Intransitivity of Preferences," Psychological Review (January 1969), pp. 31–48.

Tversky, A., and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," Science (September 1974), pp. 1124–31.

von Neumann, J., and O. Morgenstern, Theory of Games and Economic Behavior, Princeton, Ill.: Princeton University Press, 1947.

Wadsworth, G.P., and J.G. Bryan, Introduction to Probability and Random Variables, New York: MacGraw–Hill, 1960.

Wainer, H., and D. Thissen, "Graphical Data Analysis," Annual Review of Psychology (1981), pp. 191–242.

Waller, W.S., and W.L. Felix, Jr., "The Auditor and Learning from Experience: Some Conjectures," Accounting, Organizations, and Society (Fall/Winter 1984), pp. 383–406.

Wallsten, T.S., D.V. Budescu, A. Rappoport, R. Zwick, and B. Forsyth, "Measuring the Vague Meaning of Probability Terms," Journal of Experimental Psychology: General (December 1986), pp. 348–64.

Weber, R., "Some Characteristics of the Free Recall of Computer Controls by EDP Auditors," Journal of Accounting Research (Spring 1980), pp. 214–41.

Winkler, R.L., "Probabilistic Prediction: Some Experimental Results," Journal of the American Statistical Association (December 1971), pp. 675–85.

Winkler, R.L., "The Assessment of Prior Distributions in Bayesian Analysis," Journal of the American Statistical Association (September 1967), pp. 776-800 (a).

Winkler, R.L., "The Quantification of Judgment: Some Methodological Suggestions," Journal of the American Statistical Association (December 1967), pp. 1105-20 (b).

Winkler, R.L., and A.H. Murphy, "Good Probability Assessors," Journal of Applied Meteorology (October 1968), pp. 751-58.

Wright, A., "An Investigation of the Engagement Evaluation Process for Staff Auditors," Journal of Accounting Research (Spring 1982), pp. 227-39.

Wright, W.F., "Empirical Comparison of Subjective Probability Elicitation Methods," Contemporary Accounting Research (Fall 1986), pp. 47-57.

Wright, W.F., "Accuracy of Subjective Probabilities for a Financial Variable," in Burns, T.J. (ed.), Behavioral Experiments in Accounting II, Columbus, Ohio: Ohio State University, 1979, pp. 1-13.

Yeh, R.Z., Modern Probability Theory, New York: Harper and Row, 1973.

Zehna, P.W., Finite Probability, Boston, Mass.: Allyn and Bacon, 1969.

APPENDICES

page

APPENDIX A – Selected Models of Risk 95

APPENDIX B – Pre–experimental Instructions

    B.1 – Pre–experimental Instructions for Linear Scale Tutored Condition (C_{lst}) 96
    B.2 – Pre–experimental Instructions for Linear Scale Untutored Condition (C_{lsnt}) 100
    B.3 – Pre–experimental Instructions for Pie Chart Tutored Condition (C_{pct}) ....
    B.4 – Pre–experimental Instructions for Pie Chart Untutored Condition (C_{pcnt}) 108
APPENDIX C – Training Exercise 112

APPENDIX D – The Fraud Case

    D.1 – Initial Narrative 119
    D.2 – Additional Evidence 124
    D.3 – Response Format for Linear Scale Conditions (C_{lst}; C_{lsnt}) 126
    D.4 – Response Format for Pie Chart Conditions (C_{pct}; C_{pcnt}) 128
    D.5 – Confidence in Final Probability Judgments – All Conditions 129
    D.6 – Manipulation Checks 130

APPENDIX E – The Internal Control Case

    E.1 – Initial Narrative 134
    E.2 – Additional Evidence 140
    E.3 – Response Format for Tutored Linear Scale Condition (C_{lst}) 141
    E.4 – Response Format for Tutored Pie Chart Condition (C_{pct}) 142
    E.5 – Response Format for Non–Probabilistic Scale Condition (C_{nps}) 143
    E.6 – Final Choice of Level of Substantive Testing – All Conditions 144
    E.7 – Manipulation Checks 145
APPENDIX F – Demographic Information 148

page

APPENDIX G - Statistical Results
G.1 - H_4 150
G.2 - H_5 151
G.3 - Longitudinal Analysis 152
G.4 - H_7 154
G.5 - H_8 155
G.6 - H_9 156

APPENDIX A

Selected Models of Risk in Psychology and Business

TYPE	FORMULATION	PARAMETERS
Multiplicative Probability		
Expectation models		
Expected value (Pascal, Cardose)	$EV = \Sigma p_i x_i$	p_i = probability of ith outcome x_i = objective value of ith outcome outcome
Expected utility (von Neumann & Morgenstern 1947)	$EU = \Sigma p_i U(x_i)$	p_i = probability of ith outcome $U(x_i)$ = utility of the ith outcome
Subjective expected utility (Savage 1954; Edwards 1955)	$SEU = \Sigma f(p_i)U(x_i)$	$f(p_i)$ = subjective probability of the ith outcome $U(x_i)$ = utility of the ith outcome
Prospect theory (Kahneman & Tversky 1979)	$V = \Sigma\ \pi(p_i)v(x_i)$	$\pi(p_i)$ = decision weight attached to the ith outcome $v(x_i)$ = "value" of the ith outcome
Moment models		
Probability of below-target returns (Fishburn 1977)	$R(A) = \int_{-\infty}^{x} (t-x)\alpha dF(x)$	$\int_{-\infty}^{x}$ = area below the target outcome t = the actual outcome x = the target outcome α = the decision makers' attitude towards deviations from the target return $F(x)$ = the cumulative distribution function of X
Mean-variance (Markowitz 1952)	$R = U(\mu,\sigma^2)$	μ = mean or expected value of the distribution of returns σ^2 = variance of the distribution of returns
Confidence limit (Baumol 1963)	$R = \mu - k\sigma$	μ = mean or expected value of the distribution of returns k = a measure of the decision makers' tolerance towards the spread of the distribution of returns σ = standard deviation of the distribution of returns
Probability of ruin (Agnew et al. 1969)	$R = P(r_i x_i < t) \leq \alpha$	r_i = rate of return on asset i x_i = purchase price of asset i t = minimum acceptable return α = maximum acceptable probability
Additive Probability		
Risk dimensions model (Slovic 1967)	$A(G) = \alpha + P_W + \$_W + P_L + \$_L$	α = a measure of the decision makers' propensity to gamble P_W = probability of winning $\$_W$ = amount to win P_L = probability of losing $\$_L$ = amount to lose
Cumulative Probability		
Distributional model (Lopes 1984)	$R = 1-[\Sigma(Y_i/\bar{Y})^{1-\epsilon} f_i]^{1/(1-\epsilon)}$ * *This specific formulation is only a preliminary suggestion for modelling risk -- it is Lopes' adaptation of the wealth distribution model proposed by Atkinson [1970].	Y_i = amount of the ith outcome $\bar{Y}$ = mean outcome ϵ = a measure of the importance of the inequality of outcomes f_i = proportion of total value in the ith outcome

APPENDIX B.1

Pre-Experimental Instructions for Linear Scale Tutored Condition (C_{lst})

In everyday life, we often need to talk about varying degrees of certainty in our beliefs, and we use such phrases as "possible", "probable", "likely", "sure thing" and "no chance". Scientists need a more precise way of describing degrees of certainty, and one way they achieve this is through the use of PROBABILITIES. In the following two cases you will be asked to use probabilities to indicate your beliefs. Before we beging though, we need to be sure that everyone understands the rating scale and to review a few basic rules of probability.

[see page of five rating scales]

These are the scales we will be using to indicate probabilities. You'll notice that they begin at 0.0 and goes up to 1.0.

The first probability rule is: probabilities must lie somewhere in the range from zero to one. If you say that some event has a probability of zero, you are saying that it can NEVER happen. If you say that an event has a probability of one, you are saying that it is CERTAIN to happen. If you say that an event has a probability of .5, you are saying that it is just as likely to happen as it is not to happen.

For example, suppose I toss a coin. Assuming it's a fair coin, what do you think is the probability that it will land with "heads" showing? We would mark the scale so that it was evenly divided into two parts – the coin is just as likely to land with "heads" showing as not. [Tutor and subjects mark scale]

Good. Now since the probability of heads is not one, we know that it is not a sure thing. The other thing that could happen is that when we flip the coin it could come up "tails". We could use a similar scale to indicate the probability that the coin will land with "tails" showing. [Tutor and subjects mark scale] Could it ever be heads and tails at the same time? [Subjects answer no]

Clearly not. What we have is a set of two possible events where only ONE AT A TIME can occur. We say that heads and tails are MUTUALLY EXCLUSIVE.

Could we ever get NEITHER heads nor tails? [Subject answers no] We might imagine some strange situation where the coin could wind up standing on edge, but to all intents and purposes we can say that heads and tails are the only two possibilities that can occur, and thus ONE of them HAS to happen. We say that heads and tails are an EXHAUSTIVE set of outcomes because one or the other MUST happen.

Now we come to our next probability rule. When we have an EXHAUSTIVE set of MUTUALLY EXCLUSIVE outcomes, the probabilities of those outcomes must add up to one. We can see this in our example of flipping the coin – the probabilities for heads and tails, .5 and .5, add up to one. This is because they can't both happen at once and one of them has to happen. The coin flipping is a pretty trivial example, but the same thing would hold true for any number of exhaustive, mutually exclusive events.

Let's look at another example. Suppose that there are only three things that can happen in the 1988 U.S. Presidential election – either a Republican can win, or a Democrat can win, or "someone else" can win. (By "someone else" we mean either an Independent or someone from a different political party. This is to be sure that we have exhausted all the possible outcomes in a situation like this.) In this election, then, only one of these events can happen, but one of them will definitely happen. What do you think is the probability that a Republican will win? [Subject marks scale] That a Democrat will win? [Subject marks scale] That someone else will win? [Subject marks scale]

Now look at the three ratings you have made. They should add up to one – the three alternatives make up a set of EXHAUSTIVE, MUTUALLY EXCLUSIVE outcomes. Maybe you would like to go back and change one or more of the ratings so that they do sum to one?

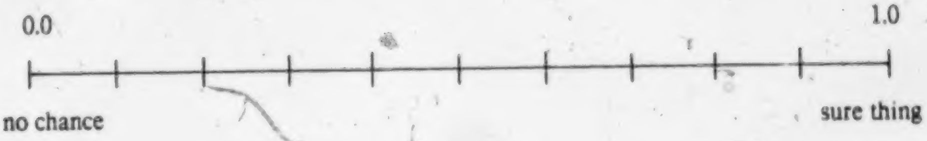
Now we're going to work through an exercise together to ensure that everyone understands how this is done.

[Training exercise - Appendix C]

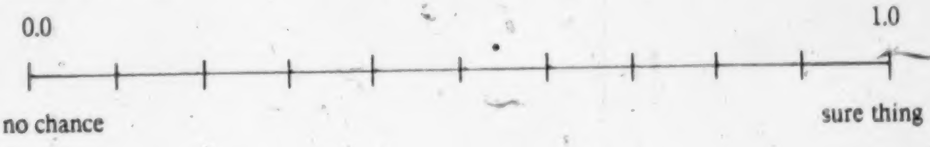
Good. Now I'm going to give you two short cases. The first one concerns a material fraud with five suspects. I want you to use the scales to indicate how likely you feel it is that the various suspects are guilty.

The second case will be a short case involving an internal control risk evaluation with four different levels of control risk. [Describe alternate levels] You will be asked to use the graphs to indicate how likely you feel it is that the internal control system is characterized by each one of the four levels risk. For example, if you think, based on the description, that the risk in this internal control system is likely to be high, you would choose a probability for that category close to 1.0. If you think that the risk of the system is extremely low, you would choose a probability for that category close to 1.0. But remember - we are going to assume that the internal control system can be characterized by those risk categories and that those four categories describe the entire range of possible risk levels.

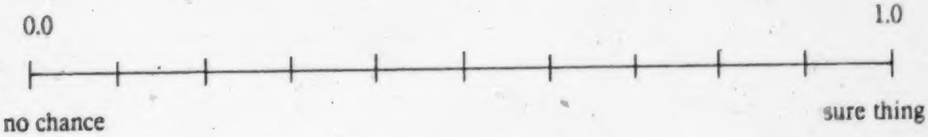
1. Probability of "heads"


0.0
1.0
no chance
sure thing

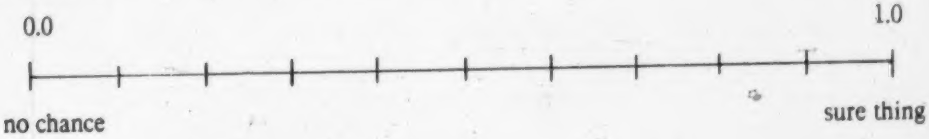
2. Probability of "tails"


0.0
1.0
no chance
sure thing

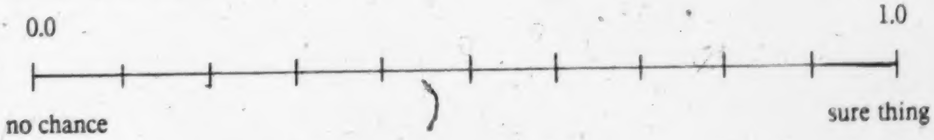
3a. Probability of "Republican"


0.0
1.0
no chance
sure thing

3b. Probability of "Democrat"


0.0
1.0
no chance
sure thing

3c. Probability of "Other"


0.0
1.0
no chance
sure thing

APPENDIX B.2

Pre-experimental Instructions for Linear Scale Untutored Condition (C_{lsnt})

In everyday life, we often want to talk about varying degrees of certainty in our beliefs, and we use such phrases as "possible", "probable", "sure thing", and "no chance". Scientists need a more precise way of describing degrees of certainty, and one way they do this is through the use of probabilities. Probability is usually represented by numbers between zero and one: a "0" probability would indicate that you believe an event has NO CHANCE of occurring, and a probability of "1" would indicate that you believe the event is ABSOLUTELY SURE to occur.

In one of the following two cases I will be asking you to use probabilities to indicate your beliefs. Before we begin, though, we should be sure that everyone understands the rating scale. Here are some examples:

1. Suppose I have a coin, and I flip it, Assuming it's a fair coin, what do you think is the probability that the coin will come come up "heads"? We would mark that portion of the graph that corresponds to the probability of heads. [Suggest .5]

2. Now suppose that I roll an ordinary 6-sided die. How would we show on the graph how likely we feel that a "2" is to be showing? [Suggest .15+]

3. We also use probabilities to describe other events which may or may not happen in the future. For instance, what do you believe is the probability that there will be rain in Vancouver tomorrow? [Mark in the scale; discuss 0 probability]

4. What do you feel is the probability that there will still be a majority Conservative government after the next federal election? [Mark in the scale]

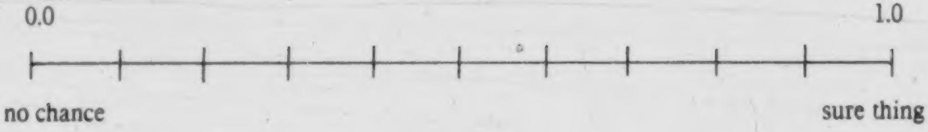
Now we're going to work through an exercise together to ensure that everyone understands how this will be done in the actual case.

[Training Exercise – Appendix C]

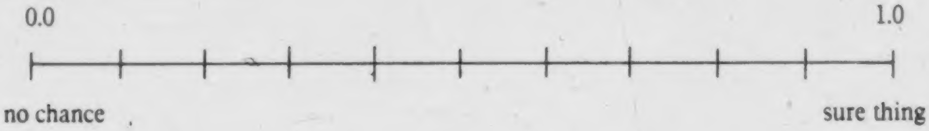
Now I'm going to give you two short cases. The first concerns a material fraud with five suspects. I want you to use the graphs to indicate how likely you feel it is that the various suspects are guilty.

The second case is an internal control evaluation. You will not have to make probability estimates – the description of the case will tell you what to do instead.

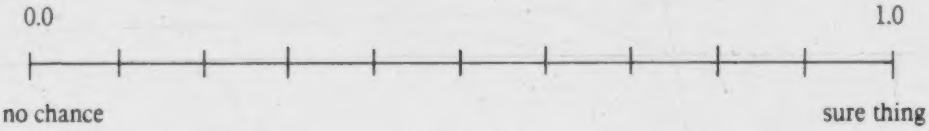
1. Probability of "heads"


0.0
1.0
no chance
sure thing

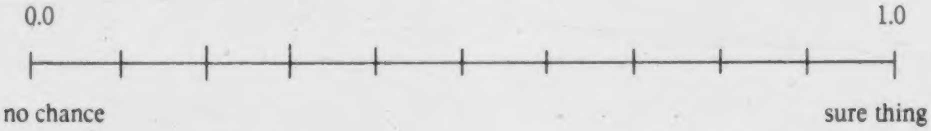
2. Probability of "2"


0.0
1.0
no chance
sure thing

3. Probability of "rain"


0.0
1.0
no chance
sure thing

4. Probability of "Conservative government"


0.0
1.0
no chance
sure thing

APPENDIX B.3

Pre-experimental Instructions for Pie Chart Tutored Condition (C_{pct})

In everyday life, we often need to talk about varying degrees of certainty in our beliefs, and we use such phrases as "possible", "probable", "likely", "sure thing" and "no chance". Scientists need a more precise way of describing degrees of certainty, and one way they achieve this is through the use of PROBABILITIES. In the following two cases you will be asked to use probabilities to indicate your beliefs. Before we beging though, we need to be sure that everyone understands the rating scale and to review a few basic rules of probability.

[see two-page handout]

On the first page of the handout are a number of pie-charts or pie graphs. The shaded segment of the graph in each example corresponds to a different probability value.

[see page one of the handout]

Following are a few simple examples to illustrate how to use the graphs.

[see page two of the handout]

The first probability rule is: probabilities must lie somewhere in the range from zero to one. If you say that some event has a probability of zero, you are saying that it can NEVER happen. If you say that an event has a probability of one, you are saying that it is CERTAIN to happen. If you say that an event has a probability of .5, you are saying that it is just as likely to happen as it is not to happen.

For example, suppose I toss a coin. Assuming it's a fair coin, what do you think is the probability that it will land with "heads" showing? We would mark the graph so that it was evenly divided into two parts - the coin is just as likely to land with "heads" showing as not. [Tutor and subjects mark graph]

Good. Now since the probability of heads is not one, we know that it is not a sure thing. The other thing that could happen is that when we flip the coin it could come up "tails". We could use a similar scale to indicate the probability that the coin will land with "tails" showing. [Tutor and subjects mark graph] Could it ever be heads and tails at the same time? [Subjects answer no] Clearly not. What we have is a set of two possible events where only ONE AT A TIME can occur. We say that heads and tails are MUTUALLY EXCLUSIVE.

Could we ever get NEITHER heads nor tails? [Subject answers no] We might imagine some strange situation where the coin could wind up standing on edge, but to all intents and purposes we can say that heads and tails are the only two possibilities that can occur, and thus ONE of them HAS to happen. We say that heads and tails are an EXHAUSTIVE set of outcomes because one or the other MUST happen.

Now we come to our next probability rule. When we have an EXHAUSTIVE set of MUTUALLY EXCLUSIVE outcomes, the probabilities of those outcomes must add up to one. We can see this in our example of flipping the coin – the probabilities for heads and tails, .5 and .5, add up to one. This is because they can't both happen at once and one of them has to happen. The coin flipping is a pretty trivial example, but the same thing would hold true for any number of exhaustive, mutually exclusive events.

Let's look at another example. Suppose that there are only three things that can happen in the 1988 U.S. Presidential election – either a Republican can win, or a Democrat can win, or "someone else" can win. (By "someone else" we mean either an Independent or someone from a different political party. This is to be sure that we have exhausted all the possible outcomes in a situation like this.) In this election, then, only one of these events can happen, but one of them will definitely happen. What do you think is the probability that a Republican will win? [Subject marks graph] That a Democrat will win? [Subject marks graph] That someone else will win? [Subject marks graph]

Now look at the three ratings you have made. They should add up to one - the three alternatives make up a set of EXHAUSTIVE, MUTUALLY EXCLUSIVE outcomes.

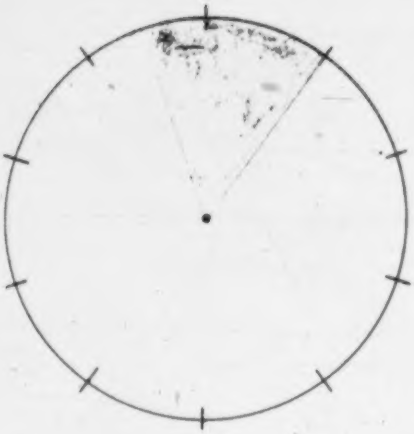
Now we're going to work through an exercise together to ensure that everyone understands how this is done.

[Training exercise - Appendix C]

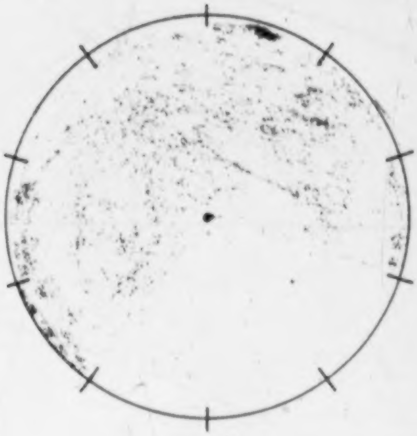
Good. Now I'm going to give you two short cases. The first one concerns a material fraud with five suspects. I want you to use the scales to indicate how likely you feel it is that the various suspects are guilty.

The second case will be a short case involving an internal control risk evaluation with four different levels of control risk. [Describe alternate levels] You will be asked to use the graphs to indicate how likely you feel it is that the internal control system is characterized by each one of the four levels risk. For example, if you think, based on the description, that the risk in this internal control system is likely to be high, you would choose a probability for that category close to 1.0. If you think that the risk of the system is extremely low, you would choose a probability for that category close to 1.0. But remember - we are going to assume that the internal control system can be characterized by those risk categories and that those four categories describe the entire range of possible risk levels.

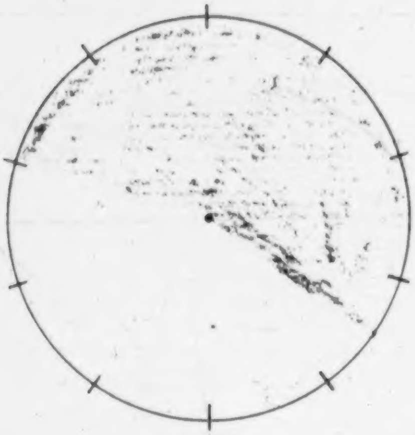
15%



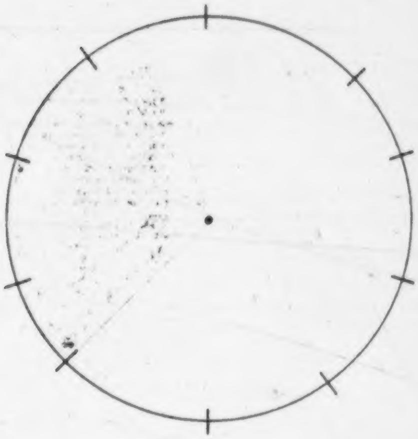
70%



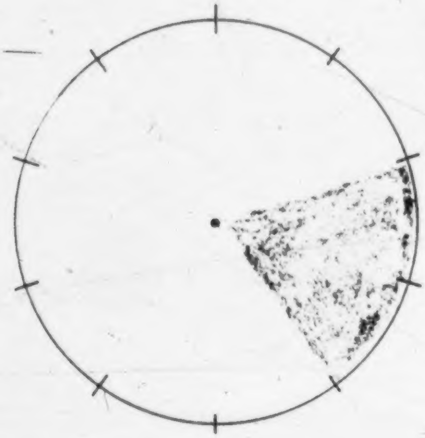
55%



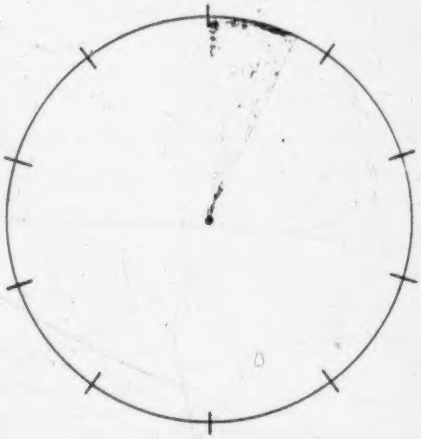
35%



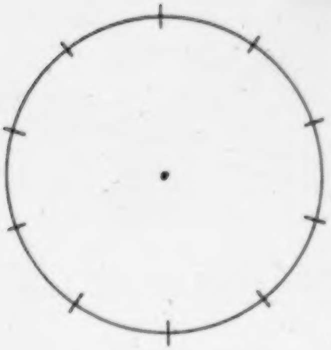
20%



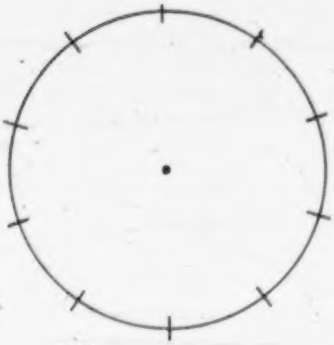
7%



1. Probability of "heads"



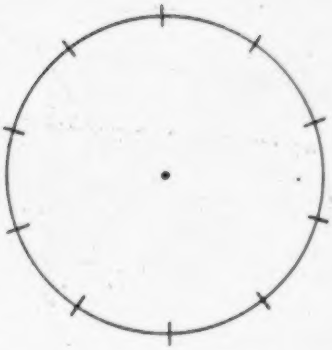
2. Probability of "tails"



3a. Probability of "Republican"

3b. Probability of "Democrat"

3c. Probability of "Other"



APPENDIX B.4

Pre-experimental Instructions for Pie Chart Untutored Condition (C_{pcnt})

In everyday life, we often want to talk about varying degrees of certainty in our beliefs, and we use such phrases as "possible", "probable", "sure thing", and "no chance". Scientists need a more precise way of describing degrees of certainty, and one way they do this is through the use of probabilities. Probability is usually represented by numbers between zero and one: a "0" probability would indicate that you believe an event has NO CHANCE of occurring, and a probability of "1" would indicate that you believe the event is ABSOLUTELY SURE to occur.

In one of the following two cases, I will be asking you to use probabilities to indicate your beliefs. Before we begin, though, we should be sure that everyone understands the rating scale. Here are some examples:

[see two-page handout]

On the first page of the handout are a number of pie-charts or pie graphs.The shaded segments of the graphs in each of the examples corresponds to the probability value printed above it.

[refer to page one of the handout]

Now we'll do a few questions to practice using the graphs.

[refer to page two of the handout]

1. Suppose I have a coin, and I flip it. Assuming it's a fair coin, what do you think is the probability that the coin will come come up "heads"?. We would mark that portion of the graph that corresponds to the probability of heads. [Suggest .5]

2. Now suppose that I roll an ordinary 6–sided die. How would we show on the graph how likely we feel that a "2" is to be showing? [Suggest .1+]

3. We also use probabilities to describe other events which may or may not happen in the future. For instance, what do you believe is the probability that there will be rain in Vancouver tomorrow? [Mark in the graph; discuss 0 probability]

4. What do you feel is the probability that there will still be a majority Conservative government after the next federal election? [Mark in the graph]

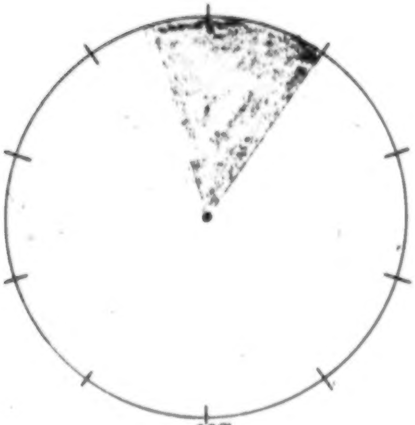
Now we're going to work through an exercise together to ensure that everyone understands how this will be done in the actual cases.

[Training Exercise – Appendix C]

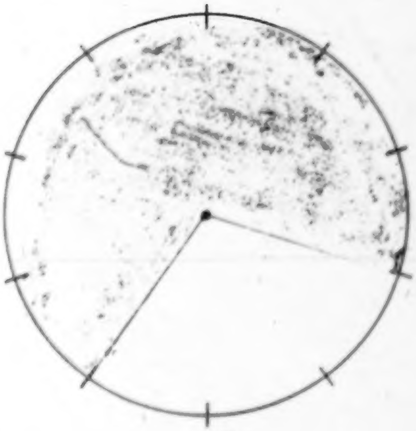
Now I'm going to give you two short cases. The first concerns a material fraud with five suspects. I want you to use the graphs to indicate how likely you feel it is that the various suspects are guilty.

The second case is an internal control evaluation. You will not have to make probability estimates – the description of the case will tell you what to do instead.

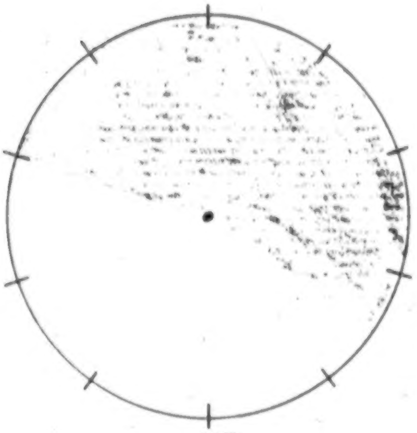
15%



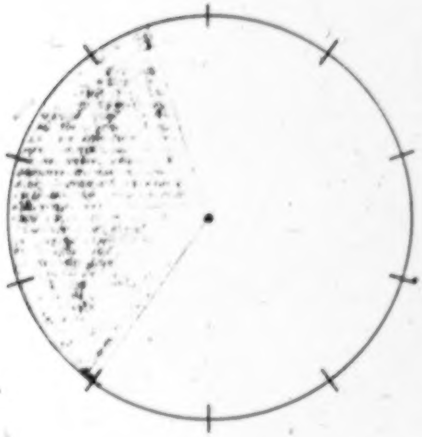
70%



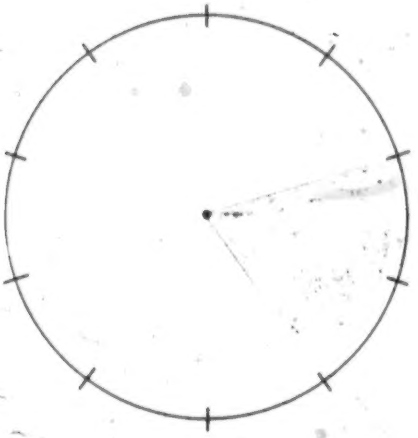
55%



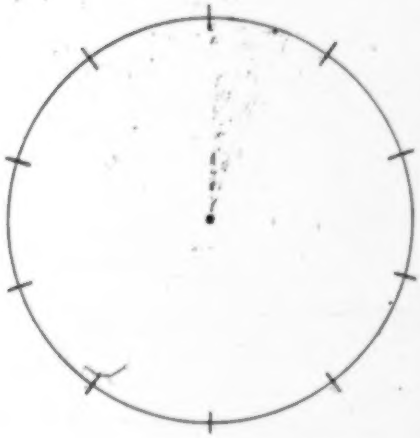
35%



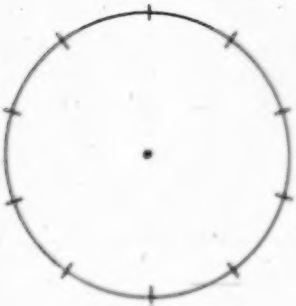
20%



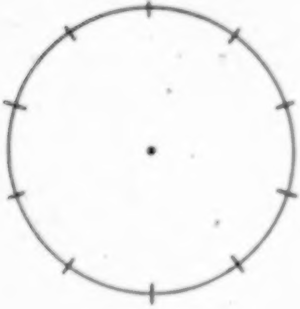
7%



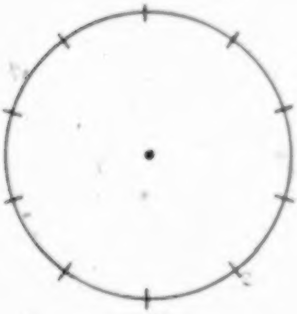
1. Probability of "heads"



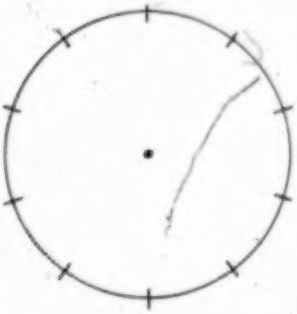
2. Probability of "2"



3. Probability of "rain"



4. Probability of "Conservative government"



APPENDIX C

TRAINING EXERCISE

Now, in order to give you a general idea of how the actual cases will proceed, we will go through a brief exercise using a murder mystery as an example...

A young high school teacher is found dead in the teachers' lounge at about 7:30 a.m. on Wednesday morning. It appears that he has been murdered – struck in the head with a heavy object. The medical examiner places the time of death at between 7:00 p.m. and 9:00 p.m. the previous evening.

In the course of the investigation it becomes evident that only the following persons can possibly have been in the vicinity of the lounge after school hours on Tuesday. By implication, one of them has to be the murderer, and the police are certain that there was only one person involved. The four suspects are:

1. The cleaning woman, Alice Wilson, who found the body when she entered the teachers' lounge Wednesday morning. She is a middle–aged woman who has been employed at the school for a number of years. She appears very upset by the incident and has difficulty giving the police a coherent statement.

2. The Principal, Alexander Madsen, who on Tuesday stayed in his office later than usual, by his account to make preparations for a staff meeting scheduled for the following day. He claims to have gone straight home from the office to have his dinner. He says he arrived home about 9:00 p.m.

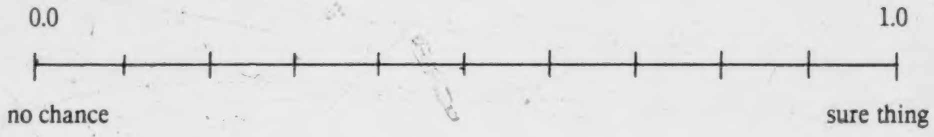
3. The Phys. Ed. teacher, Bill Kramer, who, after school hours returned to the teachers' lounge to get his sports jacket which he had left there earlier in the day. He claims to have spoken briefly with the victim, who was alive and well, at that time.

4. A student, John Beresford, who had returned after school hours with an assignment for the teacher, which he had forgotten at home that day. John says that the teacher was not in his office, as he had said he would be, and he returned home again without handing in the assignment.

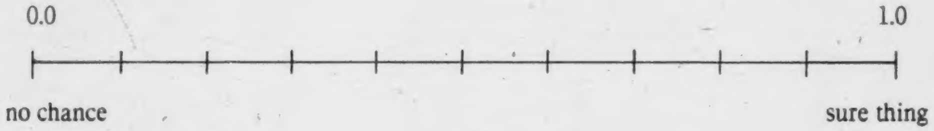
[C_{lst}; C_{lsnt}]

On the basis of this description of the four suspects, what would you say the likelihood of guilt of each one is? Please make a slash on the each of the scales below to indicate your feelings.

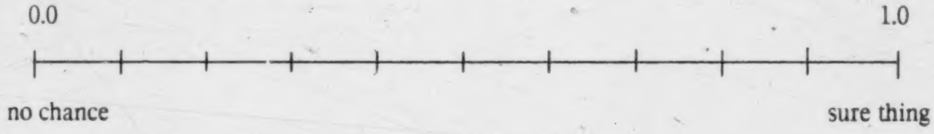
Alice Wilson (cleaning woman)


0.0
1.0
no chance
sure thing

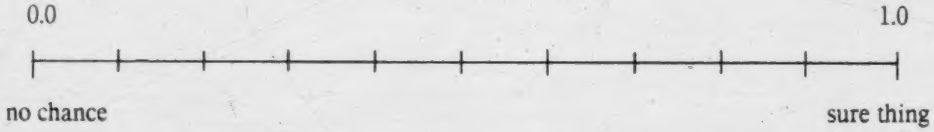
Alexander Madsen (Principal)


0.0
1.0
no chance
sure thing

Bill Kramer (Phys. Ed. teacher)


0.0
1.0
no chance
sure thing

John Beresford (student)


0.0
1.0
no chance
sure thing

[For C_{lst} emphasize that $\Sigma p_i = 1.0$]

[For C_{lsnt} just illustrate the method of response.]

[C_{pct}; C_{pcnt}]

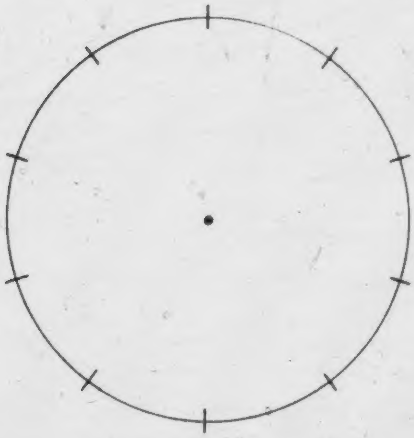
On the basis of this description of the four suspects, what would you say the likelihood of guilt of each one is? Please divide the circle below into segments to represent your feelings. Put the suspect's name or initials in the segment you have drawn to indicate that person's likelihood of guilt.

Alice Wilson (cleaning woman)

Alexander Madsen (Principal)

Bill Kramer (Phys. Ed. teacher)

John Beresford (student)



[Emphasize drawing segments and labelling segments.]

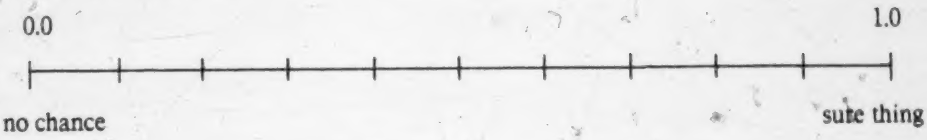
Additional Evidence

[C_{lst}; C_{lsnt}]

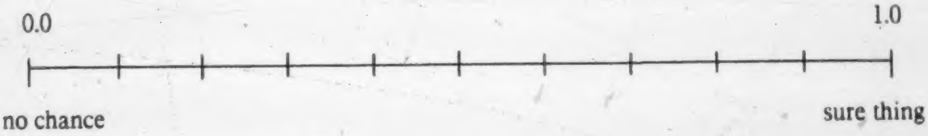
Investigation has revealed that the murder weapon was in fact a high school football trophy which normally rests in a locked case in the Phys. Ed. teacher's office. The case was not broken or forced, and the key has always been kept by the Phys. Ed. teacher who maintains the trophy display. There is also a master key to all the locks in the school but it resides in the custodians office in a locked drawer in his desk. (The custodian could not have committed the murder – he was in another town during the time in question and was seen by half-a-dozen independent witnesses.) The trophy was found, with bloodstains matching the victim's bloodtype, in a dumpster two blocks from the school.

On the basis of this additional piece of evidence, what do you feel is the likelihood of guilt of each of the suspects? Please make a slash on each of the scales below to indicate your feelings.

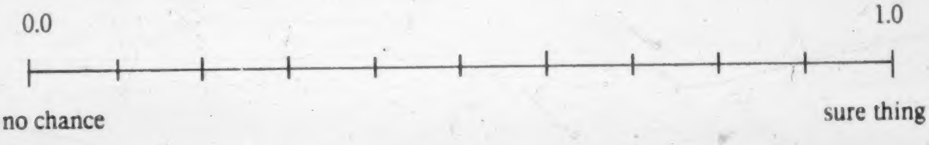
Alice Wilson (cleaning woman)


0.0
1.0
no chance
sure thing

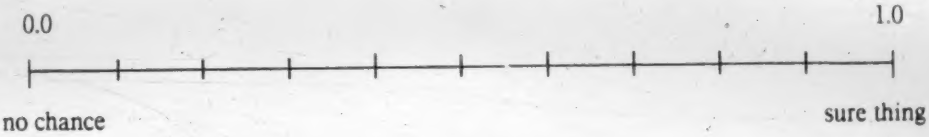
Alexander Madsen (Principal)


0.0
1.0
no chance
sure thing

Bill Kramer (Phys. Ed. teacher)


0.0
1.0
no chance
sure thing

John Beresford (student)


0.0
1.0
no chance
sure thing

[For C_{lst} emphasize that any revisions of p_i, $i = 1,...,4$, must result in $\Sigma p_i = 1.0$]

[For C_{lsnt} just indicate that it may be appropriate to make new estimates.]

Additional Evidence

$[C_{pct};C_{pcnt}]$

Investigation has revealed that the murder weapon was in fact a high school football trophy which normally rests in a locked case in the Phys. Ed. teacher's office. The case was not broken or forced, and the key has always been kept by the Phys. Ed. teacher who maintains the trophy display. There is also a master key to all the locks in the school but it resides in the custodian's office in a locked drawer in his desk. (The custodian could not have committed the murder – he was in another town at the time in question and was seen by half-a-dozen independent witnesses.) The trophy was found, with bloodstains matching the victim's blood type, in a dumpster two blocks from the school.

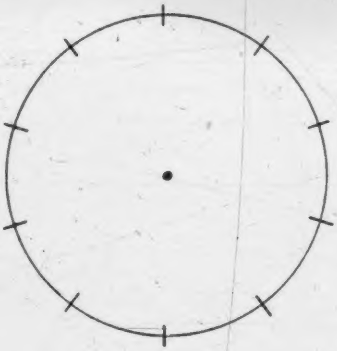
On the basis of this additional piece of evidence, what do you now feel is the likelihood of guilt of each of the suspects? ~~Please~~ divide the circle below into segments to represent your feelings. Put the suspect's name or initials in the segment you have drawn to indicate that person's likelihood of guilt.

Alice Wilson (cleaning woman)

Alexander Madsen (Principal)

Bill Kramer (Phys. Ed. teacher)

John Beresford (student)



[emphasize drawing segments and labelling segments.]

APPENDIX D.1

The Fraud Case – Initial Narrative

You are in charge of the audit of Renton Ltd., a manufacturer and supplier of parts to the automotive after-market industry (1986 sales of $50 million). Although the company is a new client for your firm, it has been in business for 25 years. It is the policy of the board of directors to change auditors every 8 years or so. The previous auditors were a well-respected national firm. Your enquiries of them have indicated that there is no professional reason not to accept the audit. The financial statements of Renton Ltd. have always received a clean audit opinion.

In the course of your preliminary review of the results of Renton Ltd. for the current year under audit, it appears that there are some discrepancies. The revenue cycle (sales-accounts receivable-cash receipts) does not appear to be consistent with the purchases cycle (purchases-accounts payable-cash disbursements). It appears that the level of purchases throughout the year has been sufficient to support a level of sales in excess of that which has been recorded. The accounts receivable and cash collections from sales are likewise lower than the amount that the level of purchases would support. A preliminary count of the inventory on hand indicates that the perpetual inventory records overstate the amount of inventory actually on hand.

You begin to suspect that some sort of fraud has been perpetrated within Renton Ltd. On one hand, it could be that some of the sales are unrecorded and the resulting assets (accounts receivable and cash receipts) have been diverted, or it could be that the discrepancy results from the fact that some of the purchased goods have been diverted and not received by the company. The amount of money involved appears to be substantial.

The audit committee of the board of directors requests that you undertake a thorough investigation of the possible fraud to determine the extent, source and perpetrator. Your first step is to identify persons within the company who could be possibly have had the opportunity to perpetrate the fraud. You are able to narrow this list down to five possible suspects, listed alphabetically on the next page:

Arthur Denton - accounting supervisor

Arthur Denton joined the company three years ago, upon completion of a Bachelor of Commerce degree. It was expected at that time that he would rise rapidly in the company. However, although he is competent at his present position, he has not really lived up to the expectations that existed when he was hired and has not advanced as anticipated.

Susan Evans - salesperson

Susan Evans joined the firm from a competitor a short time ago. She was hired with the understanding that she was to be the next sales manager within six months to a year. However, the present sales manager has decided not to take early retirement as he had originally planned and a promotion will not be forthcoming for a number of years.

John Sims - chief warehouseman

John Sims has been with the company for 12 years. He started with them as a receiving clerk. He was promoted to his present position two years ago. He is very competent in his position although his lack of formal education makes his further advancement (i.e. to management) unlikely.

William Toms - in charge of purchasing

William Toms is a long-time employee who has been with the company almost since it commenced operations. He started in the warehouse but was promoted to his present position some years ago.

Elizabeth Winslow - accounts payable supervisor

Elizabeth Winslow started with the company 15 years ago directly out of high school as an accounts payable clerk. She has been promoted several times and is now in charge of the accounts payable function. She handles this function well, but it is the feeling of management that she does not have a sufficient grasp of the overall business to advance beyond this position.

You are sure on the basis of initial investigation that the perpetrator of the fraud is one of these five people, and you also feel sure that the perpetrator acted alone - that there was no collusion among members of the company.

At this early stage what do you feel the likelihood is of each of the five suspects being the perpetrator of the fraud? Please indicate your feeling by marking a slash on the scale opposite the name of that person on the following page (C_{lst}, C_{lsnt}) /dividing the pie-graph on the following page into segments representing your likelihoods and marking each segment with the initials of the person it represents (C_{pct}, C_{pcnt}). Do not put any numerical amounts in the segments.

$[C_{lst}; C_{lsnt}]$

On the basis of the above description of the five suspects, what do you feel the likelihood of guilt of each one is?

Arthur Denton (accounting supervisor)

0.0 1.0

no chance sure thing

Susan Evans (salesperson)

0.0 1.0

no chance sure thing

John Sims (warehouseman)

0.0 1.0

no chance sure thing

William Toms (purchasing)

0.0 1.0

no chance sure thing

Elizabeth Winslow (A/P supervisor)

0.0 1.0

no chance sure thing

[C_{lst}: The sum of your likelihoods should equal 1.0]

$[C_{pct}; C_{pcnt}]$

On the basis of the above descriptions of the five suspects, what do you feel is the likelihood of guilt of each one? Please divide the circle below into segments to represent your feelings. Put the suspect's initials in the segment you have drawn to indicate your feeling about that person's likelihood of guilt.

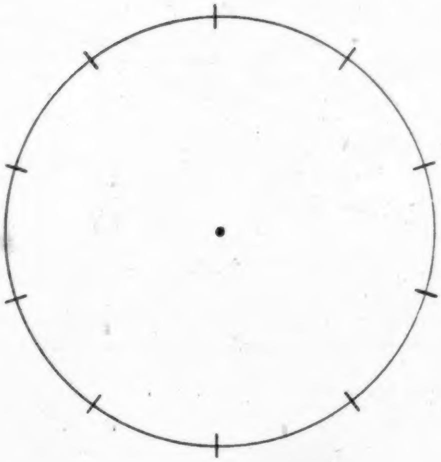
Arthur Denton (accounting supervisor)

Susan Evans (salesperson)

John Sims (warehouseman)

William Toms (purchasing)

Elizabeth Winslow (A/P supervisor)



You proceed to review the details of the company's internal control systems, and you are also able to obtain certain facts about the different suspects. Each of these pieces of evidence is presented on a separate page following. Please consider each piece of evidence in turn and indicate again your feelings about the likelihood of guilt of each of the five suspects.

APPENDIX D.2

Additional Evidence

1. In examining the executed purchase orders file you find that although they all bear evidence of proper authorization, a number of them are not supported by warehouse receiving documents. William Toms, in charge of purchasing, says that he is sure that they are simply misfiled and that he will have one of his staff locate them for you. You offer to have one of your juniors help in the search but he declines your offer.

2. Susan Evans has recently adopted a lifestyle which her salary alone could not support. According to her personnel file she lives in the Properties, and in talking to co-workers you also learn that she travels extensively – New York, Boston, Los Angeles, Hawaii – on weekends.

3. John Sims, the chief warehouseman, appears very happy in his position. His two children have both married and left the home, and his salary, combined with his wife's income, would be sufficient to provide them with a very comfortable standard of living.

4. Arthur Denton, accounting supervisor is very unhappy that he has been passed over for promotion. He has remarked that the company lied to him – that they are paying him much less than he is worth because they hired him under false pretenses and that they may regret their "lack of good faith."

5. The missing receiving documents are given to you by William Toms. You match them to the p.o.'s and scrutinize them for any evidence of falsification. One of the two receiving clerks whose initials appear on some of the receiving documents is no longer with the company, but you are able to interview the other clerk. He substantiates that the documents bearing his initials are genuine and that the goods he signed for were received.

6. It seems that John Sims had recently accumulated some very large debts. His wife's business, producing dried flower arrangements for the upscale arts and crafts outlets, was doing so well that they opened their own retail store. The venture was a costly failure. There was even a possibility that they could lose their home to satisfy the debts accumulated as a result of this failed venture. However, John Sims was able somehow to raise the large amount of cash necessary to satisfy these debts and avoid the potential disaster.

7. You learn that Susan Evans was left the home in the Properties as part of an inheritance which also included a large cash endowment.

8. You recall that an acquaintance, an internal auditor with another client firm, to you of a technique she learned in one of her courses, that of matching the postal codes of company personnel and suppliers. You use this technique and the resulting computer match of addresses reveals that the postal code of a supplier is the same as that of Arthur Denton, accounting supervisor.

APPENDIX D.3

Response Format for Linear Scale Conditions (C_{lst}; C_{lsnt})

[Additional pieces of evidence #1 to #8]

Based on this additional evidence regarding the fraud, what do you feel now are the likelihoods of guilt of the five suspects? Please make a slash on the scales below to indicate the likelihood of the guilty person being:

Arthur Denton (accounting supervisor)

0.0 1.0

no chance sure thing

Susan Evans (salesperson)

0.0 1.0

no chance sure thing

John Sims (warehouseman)

0.0 1.0

no chance sure thing

William Toms (purchasing)

0.0 1.0

no chance sure thing

Elizabeth Winslow (A/P supervisor)

0.0 1.0

no chance sure thing

[C_{1st}: The sum of your likelihoods should equal 1.0]

APPENDIX D.4

Response Format for Pie–Chart Conditions (C_{pct}; C_{pcnt})

[Additional pieces of evidence #1 to #8]

Based on this additional evidence regarding the fraud, what do you feel now are the likelihoods of guilt of the five suspects? Please divide the graph below into segments representing your likelihoods. Put the suspect's initials in the segment you have drawn to indicate that person's likelihood of guilt.

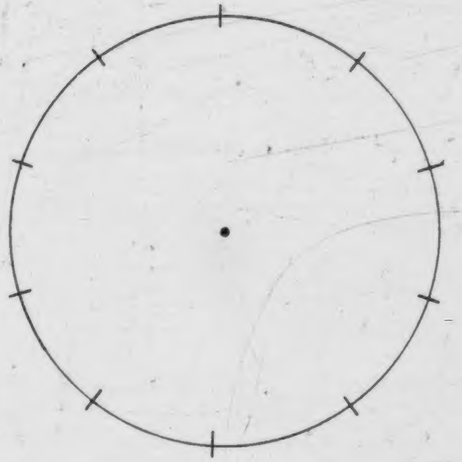
Arthur Denton (accounting supervisor)

Susan Evans (salesperson)

John Sims (warehouseman)

William Toms (purchasing)

Elizabeth Winslow (A/P supervisor)



APPENDIX D.5

Condidence in Final Probability Judgments – All Conditions

Although the investigation is not yet complete, the controller has asked you into his office, as he wishes to discuss your findings so far. Please indicate, by making a slash through the line below, how confident you are about your final overall estimate of the likelihood of guilt of the five suspects at this time. (i.e Your overall estimate after the 8th piece of evidence.)

1 7

very unconfident very confident

Please indicate, by making a slash on the line below, how difficult you felt it was to use this method of response to indicate your judgments about the probability of guilt of the suspects.

1 7

not difficult very difficult

APPENDIX D.6

Manipulation Checks

For each of the pieces of evidence presented in the foregoing case, please indicate by making a slash through the lines below, how relevant you felt that each one was in assessing the probable guilt of the suspect.

1. In examining the executed purchase orders file you find that, although they all bear evidence of proper authorization, a number of them are not supported by warehouse receiving documents. William Toms, in charge of purchasing, says that he is sure that they are simply misfiled and that he will have one of his staff locate them for you. You offer to have one of your juniors help in the search but he declines your offer.

1 7

not relevant relevant

2. Susan Evans has recently adopted a lifestyle which her salary alone could not support. According to her personnel records, she lives in the Properties, and in talking to co-workers you also learn that she travels extensively - New York, Boston, Los Angeles, Hawaii - on weekends.

1 7

not relevant relevant

3. John Sims, the chief warehouseman, appears very happy in his position. His two children have both married and left the home, and his salary, combined with his wife's income, would be sufficient to provide them with a very comfortable standard of living.

1 7

not relevant relevant

4. Arthur Denton, accounting supervisor, is very unhappy that he has been passed over for promotion. He has remarked that the company lied to him – that they are paying him much less than he is worth because they hired him under false pretenses and that they may regret their "lack of good faith".

1 7

not relevant relevant

5. The missing receiving documents are given to you by William Toms. You match them to the p.o.'s and scrutinize them for any irregularities. One of the receiving clerks whose initials appear on some of the receiving documents is no longer with the company, but you are able to interview the other receiving clerk. He substantiates that the documents bearing his initials are genuine and that the goods he signed for were received.

1 7

not relevant relevant

6. It seems that John Sims had recently accumulated some very large debts. His wife's business, producing dried flower arrangements for upscale arts and crafts outlets, was doing so well that they opened their own retail store. The venture was a costly failure. There was even a possibility that they would lose their home to satisfy the debts accumulated as a result of this failed venture. However, John Sims was able somehow to raise the large amount of cash necessary to satisfy these debts and avoid the potential disaster.

1 7

not relevant relevant

7. You learn that Susan Evans was left the home in the Properties as part of an inheritance which also included a large cash endowment.

1 7

not relevant relevant

8. You recall that an acquaintance, an internal auditor with another client, told you of a technique she learned in one of her courses, that of matching the postal codes of company personnel and suppliers. You use this technique and the resulting computer match of mailing addresses reveals the fact that the postal code of a fairly large supplier is the same as that of Arthur Denton, accounting supervisor.

1 7

not relevant relevant

Please indicate, by making a slash on the line below, how difficult you felt it was to use this method of response to indicate your judgments about the probability of guilt of the suspects.

1 ——————————————————————— 7

not difficult ——————————————— very difficult

APPENDIX E.1

The Internal Control Case - Initial Narrative,

You are the auditor in charge of the year-end audit of the Exdan Company. The purpose of your audit is to issue a standard short-form audit report. Your firm has performed the annual audit for the past several years, but this is the first time that you have been in charge of the field work. Your firm issued an unqualified opinion last year.

The Exdan Company produces air conditioners in an assembly-line operation. It is a publicly held company with about 5,000 stockholders. During the past year, net sales were about $150 million, a small increase over the previous year.

In accordance with GAAS, an investigation of the company's internal control system will be undertaken before you determine the nature, extent and timing of substantive audit procedures which should be applied. You have several assistants who will review the existing internal control system and the operation of that system in the various areas of the company's operations; for instance, one assistant will investigate the internal controls pertaining to cash receipts; another will review the controls over marketable securities; still another will review the controls over payroll, etc. You will then review the results of each assistant's work, and based on your evaluation of the internal control risk, you will prescribe the appropriate audit procedures to be applied in each area.

In this case you are concerned only with the internal controls over payroll. The company has about 400 factory employees, an increase of about 25 over the previous year. The employees are paid bi-weekly and the annual payroll is about $12 1/2 million. Hourly wage rates are established in the union contract.

When a new employee is hired by the company, the personnel department sends a copy of the TD1 and other voluntary deduction cards to the payroll department. The personnel department also notifies the payroll department when a worker's employment is terminated.

A time clock is used. The cards are kept in a rack beside the clock, and factory employees are required to punch in and out every day.

The payroll department uses the time cards as input to the payroll computation process. Other input, in the form of authorized wage rates and authorized deductions, is obtained from the personnel department. After the payroll cheque-run is completed, the cheques are sent to the controller's office along with any cheques that were spoiled in the preparation process.

The cheques are machine-signed by the controller's office and a copy of the cheque-run summary is sent to the treasurer's office where an authorization is prepared for the exact amount of the total net pay to transfer funds from the general bank account to the payroll bank account. After the treasurer signs the authorization, his secretary sends it to the bank to have the funds transferred to the payroll bank account the day before the payroll cheques are to be issued.

The people directly concerned with the payroll system have been with the company an average of six years, with the range being two to ten years.

For the purpose of responding to the questions which follow, suppose that you have instructed your assistant to examine the internal control pertaining to payroll, upon which you will base your judgment of the internal control risk. Your assistant has an internal control questionnaire which contains a list of the relevant features of a payroll internal control system. Please consider each of the characteristics of the client's internal control system in turn and, after considering each one, estimate the likelihood of the internal control risk being extremely low, low, moderate, or high. After considering the fifth piece of evidence as ascertained by your assistant you are asked to select an appropriate level of substantive testing. After choosing a level of substantive testing you will be asked to indicate how confident you are that this level is appropriate. However, based on the foregoing description, before considering the additional evidence, please indicate on the next pages your initial judgment about (C_{lst}; C_{pct}; C_{nps}: the level of risk in the payroll system and) the amount of substantive testing which is appropriate for this payroll system.

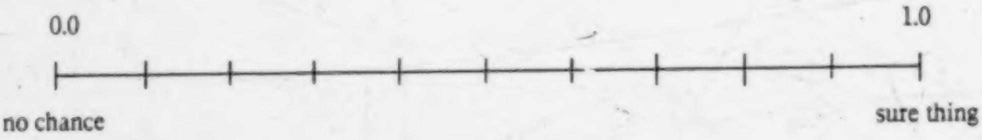
$[C_{lst}]$

Which of the following categories, in your judgment, best describes the level of risk in this internal control system? (please circle one).

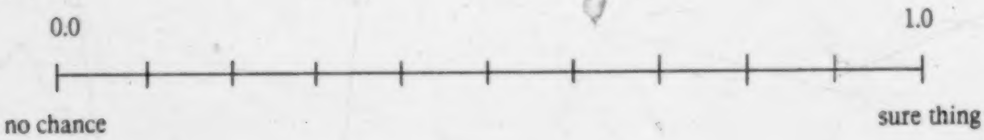
extremely low   low   moderate   high

Now please try to refine your judgment by assigning probabilities to each of the four categories above, using the scales below.

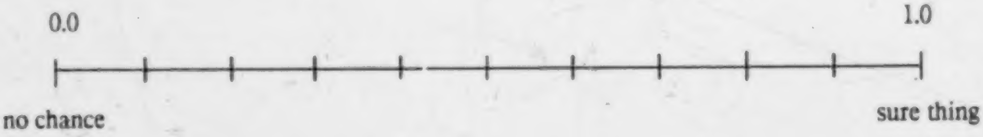
extremely low


0.0
1.0
no chance
sure thing

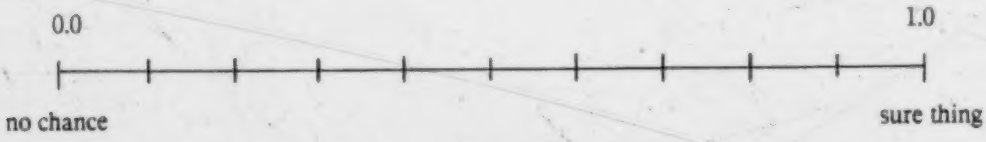
low


0.0
1.0
no chance
sure thing

moderate


0.0
1.0
no chance
sure thing

high


0.0
1.0
no chance
sure thing

[The sum of your likelihoods should equal 1.0]

[C_{pct}]

Which of the following four categories, in your judgment, best describes the level of risk in this internal control system? (please circle one)

extremely low   low   moderate   high

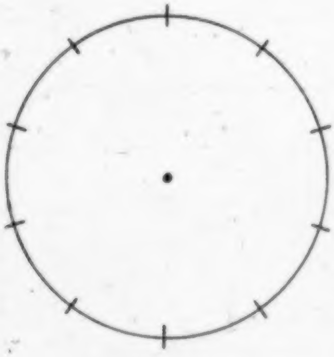
Now please try to refine your judgment by assigning probabilities to each of the four categories above using the graph below.

extremely low

low

moderate

high



[C_{nps}]

Which of the following four categories, in your judgment, best describes the level of risk in this internal control system? (please circle one)

extremely low low moderate high

Now please try to refine your judgment using the scale below.

1 7

extremely low high

[All Conditions]

On the basis of the initial description of the internal control over payroll, what level of substantive testing do you feel is appropriate for the payroll system? Please indicate your judgment by making a slash through the scale below.

The amount of substantive testing should be:

1 7

minimal exhaustive

How confident are you that the level of substantive testing you have chosen is the proper level? Please indicate your feeling by making a slash through the scale below.

1 7

very unconfident very confident

APPENDIX E.2

Additional Evidence

1. The tasks of timekeeping and payment of employees are adequately separated from the task of payroll preparation.

2. The task of payroll preparation is not adequately separated from the task of payroll bank reconciliation.

3. Formal procedures are in place for adding and deleting names on the payroll, and for making changes in pay rates and deductions.

4. The names on the payroll are not checked periodically against the active employee file of the personnel department.

5. The internal audit department has not audited payroll in the last four years.

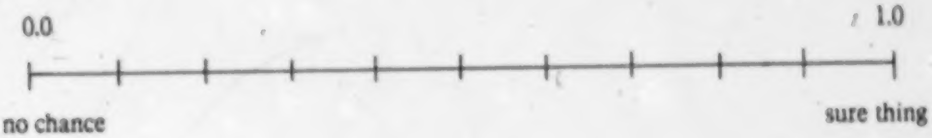
APPENDIX E.3

Response Format for Tutored Linear Scale Condition (C_{lst})

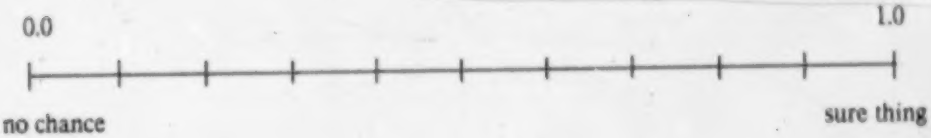
[additional pieces of evidence #1 to #5]

Based on this additional evidence regarding the internal control system, what do you feel now are the likelihoods of the four levels of risk in the payroll internal control system? Please make a slash on the scales below representing the likelihoods of the levels of internal control risk being:

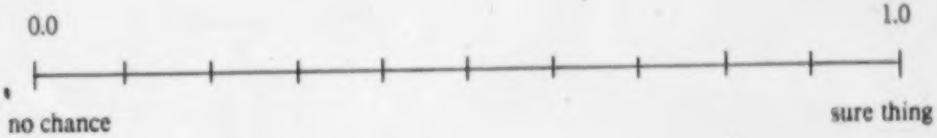
extremely low


0.0
1.0
no chance
sure thing

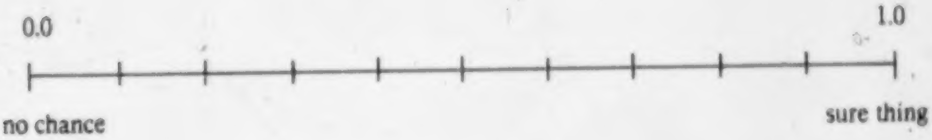
low


0.0
1.0
no chance
sure thing

moderate


0.0
1.0
no chance
sure thing

high


0.0
1.0
no chance
sure thing

[The sum of your likelihoods should equal 1.0]

APPENDIX E.4

Response format for Tutored Pie–Chart Condition (C_{pct})

[Additional pieces of evidence #1 to #5]

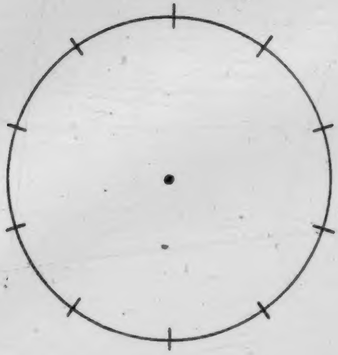
Based on this additional evidence regarding the internal control system, what do you now feel are the likelihoods of the four levels of risk in the internal control system? Please divide the graph into segments representing your likelihoods and label the segments with the initials of the category of internal control risk (e.g. "el" for extremely low, "l" for low, "m" for moderate, "h" for high).

extremely low

low

moderate

high



APPENDIX E.5

Response Format for Non-Probabilistic Scale Condition (C_{nps})

[Additional pieces of evidence #1 to #5]

Based on this additional evidence regarding the internal control system, what do you feel now is the level of risk? Please make a slash through the line below to indicate your judgment.

1 7

extremely low high

APPENDIX E.6

Final Choice of Level of Substantive Testing – All Conditions

After considering the evidence to date regarding the internal control over payroll, what level of substantive testing do you feel is appropriate for the payroll system? Please indicate your feeling by making a slash through the scale below.

The amount of substantive testing should be:

1 7

minimal exhaustive

How confident are you that the level of substantive testing you have chosen is the proper level? Please indicate your feeling by making a slash through the line below:

1 7

very unconfident very confident

APPENDIX E.7

Manipulation Checks

For each of the pieces of evidence presented in the foregoing case, please indicate by making a slash through the scales below, how relevant you felt that each one was in assessing the probability of the level of risk in the internal control system.

1. The tasks of timekeeping and payment of employees are adequately separated from the task of payroll preparation.

1 — 7

irrelevant — relevant

2. The task of payroll preparation is not adequately separated from the task of payroll bank reconciliation.

1 — 7

irrelevant — relevant

3. Formal procedures are in place for adding and deleting names on the payroll, and for making changes to pay rates and deductions.

1 7

irrelevant relevant

4. The names on the payroll are not checked periodically against the active employee file of the personnel department.

1 7

irrelevant relevant

5. The internal audit department has not audited payroll for the last four years.

1 7

irrelevant relevant

$[C_{lst}; C_{pct}; C_{nps}]$

Please indicate, by making a slash through the line below, how difficult you felt it was to use this method of response to indicate your judgments about the risk of the internal control system.

1 7

not difficult very difficult

APPENDIX F

Demographic Information

The following information is to be used only for grouping the responses for statistical analysis. There is no intention of attempting to identify you or your firm.

1. How long have you worked as an articling student/chartered accountant?____________ yrs.
(Please answer to nearest one–half year)

2. What is your present position in your firm? ______________________________________
(staff accountant, senior, supervisor,...)

3. Do you have a university degree? ____
(If you answered "no" to question 3, please skip to question 6)

4. What was your degree? ____
(Please be as specific as possible, e.g. B. Comm. (major Accounting); B.Sc. (Physics). If you have more than one degree please list them all.)

5. In what year(s) did you obtain your university degree(s)? ____________________

6. Have you had any previous training in probability theory?

_____ one or more courses in probability theory

_____ probability theory as part of another course

_____ no formal training in probability theory

(Please check one)

7. Have you ever worked on an audit where fraud was suspected? ____________________

(if the answer to question 7 is "no" , it is not necessary to answer question 8)

8. Were you involved in the investigation of the fraud? ____________________

APPENDIX G.1

Statistical Analysis of H_4

FRAUD DETECTION

(i) 2 x 2 ANOVA on "Confidence" Scores

Source	SS	DF	MS	F	Significance
Response mode	23.044	1	23.044	16.84	0.0001
Level of tutoring	5.130	1	5.130	3.75	0.0549
Interaction	4.828	1	4.828	3.53	0.0624
Error	186.059	136	1.368		

(ii) Newman-Keuls post hoc comparison of differences between means:

	C_{pct}	C_{pcnt}	C_{lsnt}
C_{pct}			
C_{pcnt}	ns @ .05		
C_{lsnt}	ns @ .05	ns @ .05	
C_{lst}	$p < .01$	$p < .01$	$p < .01$

(iii) Pairwise t-tests between means:

	C_{pct}	C_{pcnt}	C_{lsnt}
C_{pct}			
C_{pcnt}	$p = .965$		
C_{lsnt}	$p = .170$	$p = .142$	
C_{lst}	$p = .0000$	$p = .0002$	$p = .0022$

APPENDIX G.2

Statistical Analysis of H_5

FRAUD DETECTION

(i) 2 x 2 ANOVA on Difficulty Scores

Source	SS	DF	MS	F	Significance
Response mode	42.570	1	42.570	23.46	0.0000
Level of tutoring	8.064	1	8.064	4.44	0.0369
Interaction	11.315	1	11.315	5.23	0.0137
Error	246.800	136	1.815		
One-way ANOVA with variances not assumed to be equal					
Welch		3,74		16.69	0.0000
Brown-Forsyth		3,119		11.38	0.0000

(ii) Newman-Keuls post hoc comparison of differences between means:

	C_{pct}	C_{pcnt}	C_{lsnt}
C_{pct}			
C_{pcnt}	ns @ .05		
C_{lsnt}	ns @ .05	ns @ .05	
C_{lst}	$p < .01$	$p < .01$	$p < .01$

(iii) Pairwise t-tests between means:

	C_{pct}	C_{pcnt}	C_{lsnt}
C_{pct}			
C_{pcnt}	p = .8115		
C_{lsnt}	p = .0449	p = .1342	
C_{lst}	p = .0000	p = .0000	p = .0010

APPENDIX G.3

Statistical Results of Longitudinal Analyses

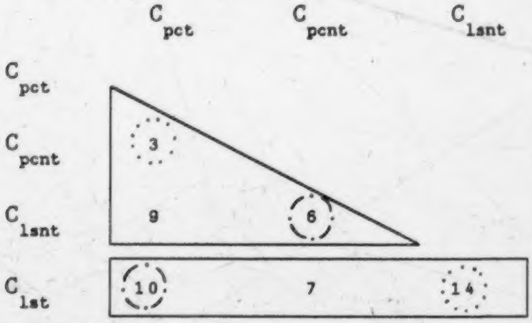
FRAUD DETECTION

(i) Profile analyses of probability of guilt estimates

Parallelism of Profiles

Suspect	All Four Treatment Conditions	Excluding C_{1st}
Arthur Denton	rejected @ $p < .001$	ns @ .05
Susan Evans	rejected @ $p < .001$	rejected @ $p < .047$
John Sims	rejected @ $p < .001$	ns @ .05
William Toms	rejected @ $p < .007$	rejected @ $p < .028$
Elizabeth Winslow	rejected @ $p < .001$	ns @ .05

(ii) Matrix of pairwise "interactions" (criss-crossing) of profiles over 9 iterations x 5 suspects



1. 18 vs. 31 difference in frequency sig. @ $p < .025$

2. 3 vs. 14 difference in frequency sig. @ $p < .001$

3. 10 vs. 6 difference in frequency ns @ .05

(iii) 2 X 2 Repeated Measures Analyses of probability of guilt estimates

						Significance	
	Source	SS	DF	MS	F	Greenhouse-Geisser	Huynh-Feldt
AD	estimate	84393.783	8	10549.223	198.30	0.000	0.000
	es	820.186	8	102.523	1.93	0.129	0.126
	* et	2700.814	8	337.602	6.35	0.005	0.004
	est	738.329	8	92.291	1.73	0.163	0.161
	error	57879.778	1088	53.198			
SE	estimate	44211.044	8	5526.381	136.20	0.000	0.000
	es	569.387	8	71.173	1.75	0.153	0.151
	et	456.657	8	57.082	1.41	0.239	0.238
	est	823.997	8	102.999	2.54	0.054	0.052
	error	44145.359	1088	40.574			
JS	estimate	37701.392	8	4712.674	91.12	0.000	0.000
	es	425.659	8	53.207	1.03	0.386	0.388
	et	765.897	8	95.737	1.85	0.129	0.125
	* est	2337.897	8	292.237	5.65	0.004	0.003
	error	56270.267	1088	51.719			
WT	estimate	54294.851	8	6786.856	159.93	0.000	0.000
	es	618.052	8	77.256	1.82	0.131	0.127
	* et	1529.133	8	191.142	4.50	0.002	0.002
	est	473.828	8	59.229	1.40	0.238	0.236
	error	46169.748	1088	42.435			
EW	estimate	4671.319	8	583.915	40.72	0.000	0.000
	es	206.651	8	25.831	1.80	0.117	0.113
	* et	396.057	8	49.507	3.45	0.006	0.005
	est	120.057	8	15.007	1.05	0.387	0.388
	error	15601.524	1088	14.400			

APPENDIX G.4

Statistical Analysis of H_7

INTERNAL CONTROL EVALUATION

(i) 4 X 1 ANOVA on "Before" scores.

Source	SS	DF	MS	F	Significance
Between	5.74	3	1.911	1.652	.1804
Within	157.36	136	1.571		
Total	163.10	139			

(ii) Non-parametric tests for equality of variances

(a) test statistic: $T_2 = \left[\sum_{j=1}^{k} \frac{S_j^2}{n_j} - N(\bar{S})^2\right] \frac{1}{D^2}$ with k-1 d.f.

$$T_2 = 8.71, \ p < .05$$

(b) multiple comparisons

	C_{pct}	C_{1st}	C_{nps}
C_{pct}			
C_{1st}	ns @ .05		
C_{nps}	ns @ .05	ns @ .05	
C	ns @ .05	$p < .05$	$p < .05$

(iii) 4 X 1 ANOVA on "After" scores.

Source	SS	DF	MS	F	Significance
Between	18.09	3	6.028	4.607	.0042
Within	177.96	136	1.309		
Total	196.05	139			

(iv) Newman-Keuls post hoc comparisons of differences between means

	C_{pct}	C_{1st}	C_{nps}
C_{pct}			
C_{1st}	ns @ .05		
C_{nps}	$p < .05$	$p < .05$	
C	$p < .05$	$p < .05$	ns @ .05

APPENDIX G.5

Statistical Tests of H_8

INTERNAL CONTROL EVALUATION

(i) 4 X 1 ANCOVA on "Confidence" scores

Source	SS	DF	MS	F	Significance
Between means	4.175	3		214.31	.0000
Covariates	90.122	1	90.122		
Error	56.770	135	.42052		
Regression	89.003	1			
Equal adj. means	5.293	3	1.764	4.196	.0071
Error	56.770	135	.42052		
Total	151.07	139			

(ii) Paired comparisons of means

	C_{pct}	C_{1st}	C_{nps}
C_{pct}			
C_{1st}	$p < .0067$		
C_{nps}	$p < .0519$	ns @ .05	
C	$p < .0013$	ns @ .05	ns @ .05

APPENDIX G.6

Statistical Tests of H_9

INTERNAL CONTROL EVALUATION

(i) 3 X 1 ANCOVA on "Difficulty" scores

Source	SS	DF	MS	F	Significance
Between	47.04	2	23.52	18.088	.0000
Within	132.64	102	1.300		
Total	179.68	104			

(ii) Newman-Keuls post hoc comparison of differences between means

	C_{pct}	C_{nps}
C_{pct}		
C_{nps}	ns @ .05	
C_{1st}	$p < .01$	$p < .01$